MENTAL MEASUREMENT MONOGRAPHS

SERIAL No. 1

AN ANALYSIS OF LANGUAGE FACTORS IN INTELLIGENCE TESTS

DOROTHY W. SEAGO, PH.D.
Hollins College, Hollins, Va.

Published by
THE WILLIAMS & WILKINS COMPANY
BALTIMORE, MD., U. S. A.

MENTAL MEASUREMENT MONOGRAPHS

This series has been established to fill the pressing need for the publication of the results of research on problems of individual psychology and the closely related problems of age, sex, and race difference.

MENTAL MEASUREMENT MONOGRAPHS is a periodical, published at irregular intervals. 400 pages to the volume. Price $5.00. (Foreign, $5.50.)

This series is intended for persons interested in individual and applied psychology, sociology, education, and mental hygiene.

The Editors are Dr. Knight Dunlap, Johns Hopkins University; and Dr. Buford Johnson, Johns Hopkins University. The Associate Editors are John E. Anderson, Yale University; John E. Coover, Stanford University; Donald G. Paterson, University of Minnesota; Joseph Peterson, George Peabody College for Teachers.

AN ANALYSIS OF LANGUAGE FACTORS IN INTELLIGENCE TESTS

DOROTHY W. SEAGO, PH.D.

Hollins College, Hollins, Va.

CHAPTER I

CHAPTER II

CHAPTER III

CHAPTER IV

CHAPTER I

Introductory and Historical

I. THE PROBLEM

The purpose of this study is to ascertain, if possible, the nature of language factors and the extent to which they function in intelligence tests; whether a facility in the understanding and use of language, as such, conditions success in intelligence tests to as great or a greater degree than other factors; the extent to which the language factor in intelligence tests makes them predictive of specific abilities.

The historical account gives a general survey of tests seeming to measure language ability directly or indirectly, followed by a more detailed study of the opposites, completion, vocabulary, information and analogies tests, as to their origin, development, uses and relationships.

In the experimental section, three representative intelligence examinations have been analyzed: (*a*) two forms of the "Thurstone psychological examination for college freshmen and high school seniors"; (*b*) the "Anderson psychological examination"; and (*c*) the "Johns Hopkins combination test." The method of correlation and partial correlation has been employed in an effort to determine the interrelation of the sub-tests of the three examinations and to isolate factors common to certain tests. Grades in specific subjects and average grades for students grouped into engineering students and arts and sciences students have been used as criteria by which to measure the specificity of given types of tests. An analysis of the distribution of grades used in this study has been included.

It will become apparent that it has not been possible in every case to find definite objective evidence that certain tests are related to each other because of a common language factor, but careful examination of the nature of the test material, together with such objective data as are available seems to justify the conclusion that they are.

Finally, one specific test, the synonym-antonym test, which has given evidence of being an especially effective single measure of a type of ability that might well be called language ability, as opposed to the ability to handle mathematics and the mathematical sciences, has been analyzed in detail, as to the types of distributions that it yields with groups of college students, and the nature of response to and the relative difficulty of specific stimulus words. A criticism of the method of scoring is made, based on the results of repetition of the test with the same subjects. Finally, a new form in which the test may be given is presented together with data as to the variability of responses in the old and new forms of the test.

II. TESTS OF LANGUAGE ABILITY

The study of language is enormously complex, and a study of the psychological aspects of language is hardly less complex, which is a necessary correlate of the importance of language in both perceptual and thought reactions.

Without becoming involved in the controversy over the immediate relation of thinking to the language mechanism, and without implying that the only medium of thought is language, it can be said that language, more specifically verbal language, is the chief medium of thought, and that, therefore, it is of value to study growth and facility in the use of language as an index to growth and facility in perceptual and thought processes.

Many investigators, among them Terman (51) and Kirkpatrick (31) have emphasized the fact that language development "mirrors" the entire mental development and that the mastery of language is a reliable index of intellectual development.

All of the methods devised for measuring the state of mental development, with the exception of a few performance tests, demand an understanding of and ability to use verbal language, but, in general, the more useful methods are those which involve the comprehension and use of verbal language to a greater degree. The extent to which facility in language is the main factor conditioning success or failure in a test is difficult to determine. Any attempt to analyze, *a priori*, just what a test is measuring is fraught with difficulties, and probably, no one analyst will

agree entirely to the analysis of another. It is true, however, that nearly all tests may be characterized by some outstanding factor.

A search through the mental test literature for data pertaining to the measurement of "language ability" reveals the fact that on one hand there are language tests specifically designated as such but also variously called tests of "general intelligence," "tests of selective thinking," "association tests," such as the Trabue Completion Test. On the other hand, we find tests which though never referred to specifically as language tests have an important language factor. The opposites test is an illustration. A study of the language factors involved in intelligence tests will thus have to take account of all tests involving the use of language, whether specifically designated language tests or not.

The division of language ability into comprehension or understanding of language and construction or use of language, as suggested by Pintner and Paterson (39), seems a useful one if we recognize that they do not represent two discrete abilities but phases of the same general ability. The ability to construct language naturally presupposes an ability to understand language. But this division of language ability into two big subheads, comprehension and construction, is a first step in analysis. In a study of tests seeming to measure language ability in some way, it soon becomes apparent that some tests measure both phases, others, but one. There are many tests measuring one aspect of language, primarily. Thus, among the tests which stress comprehension or understanding of words is, first, the vocabulary test in all of its forms. The size or range of an understanding or reading vocabulary is measured rather than the vocabulary used in speaking or composition. It may be argued that the opposites tests and similar tests requiring the ability to deal with relationships such as part-whole, subject-verb, etc., are more correctly classified as measures of constructive ability in language, but the degree to which comprehension is involved in them justifies their classification in this category. This is especially true of the opposites test in the synonym-antonym form, in which no words have to be supplied, but in which a thorough understanding of the mean-

ings of the pairs of words presented is an essential preliminary to the judging of the relationships existing between them. Other tests belonging in this category are the analogies test, especially if the stimulus words are difficult, and reading tests which are closely related to vocabulary tests, together with the closely allied tests of comprehension of questions.

Although an understanding of language is a prerequisite to construction of language, this latter involves something more. You may understand a word through its associations with other words in a given context but be altogether unable to produce the word voluntarily for use in speaking or writing. Since accuracy and fine discrimination in the choice and use of words seem to be closely related to accuracy in perceptual and thought processes, it follows that some measure of the degree of accuracy should be valuable. Among the more or less adequate measures of this sort we may mention the completion test in its various forms, as mutilated text to be restored, or as disarranged sentences; tests of linguistic invention listed by Whipple (65), *i.e.*, the sentence formation tests, both the Masselon and Meumann methods; the completion of sentences test (differing from the Ebbinghaus completion); the invention of stories, and the development of a theme; the word building test, also described by Whipple; the giving of rhymes test; and, with less justice perhaps, the naming words or free association test.

This array of names of tests will connote but little to anyone unfamiliar with the tests themselves. The mere grouping of tests into these two groups means scarcely more unless some relationship can be found to exist between them, on the basis of actual experimentation. Some of the tests mentioned have attracted the attention of investigators more than others. A more or less careful and detailed study of these tests, historically considered, may throw some light upon their real nature and offer some justification for this *a priori* classification. The following tests will be studied: (*a*) opposites tests (including the synonym-antonym test), (*b*) completion tests, (*c*) vocabulary and information tests, (*d*) analogies test. Each test will be discussed as to its origin, development, use particularly as a measure of general intelligence, or in test series designed to measure general

ability, and correlation with other tests, grades and all available criteria. The discussion of the individual tests will be followed by a summary which will attempt to relate the tests to each other.

III. DETAILED STUDY OF SPECIFIC TESTS

a. The opposites test

The value of tests of ability to distinguish differences or to note similarities between two objects or between abstract ideas as represented by words has been widely asserted. Probably no other single type of test has received more attention.

In reviewing the origin, development, and uses of this type of test, it will be necessary to distinguish between three forms: the giving of differences and similarities between two objects not present to sense, or between two abstract words; the giving of opposites in response to given stimulus words; the synonym-antonym test in which the relationship between a given pair of words has to be judged. Each of the three forms will be considered separately.

Distinguishing differences and similarities. Probably the first formal attempt to measure the ability to distinguish differences and note similarities between objects not present to sense was made by Binet and Simon (5) in three tests included in their 1905 series. In test 16, entitled "Comparison of known objects from memory," which is an eight year test in the 1908 and 1911 revisions, the subject is asked to give the difference between paper and cardboard, a fly and a butterfly, a piece of wood and a piece of glass. They consider it "an exercise in ideation, in the notion of differences, and somewhat in the power of observation" and valuable for bringing into play "the natural good sense." The test requiring the giving of differences between a president and a king which appears first as an adult test in the 1911 revision is closely related to test 16, but differs from it in being more of a test of education.

In test 20, "Resemblances of several known objects given from memory," the subject is asked to note the similarity between a poppy and blood; a fly, an ant, a butterfly and a flea; a newspaper, a label, and a picture. They consider it "a test of memory,"

of conscious recognition of resemblances, of power of observation. They note that children find considerably more difficulty in seeing similarities than in distinguishing differences and offer as partial explanation that "perhaps the needs of practical life turn their attention more towards the perception of differences than of resemblances, which only become apparent in scientific studies."

In test 30, "Definitions of abstract terms," which is closely related to test 16, a greater knowledge of vocabulary is demanded in that the subject is asked to give the difference between abstract words such as esteem and affection, weariness and sadness. They find this a difficult test for normal children of twelve years and one that differentiates clearly between subnormals and normals. With some change in stimulus words it is placed as a thirteen year test in the 1908 revision and as an "adult" test in the 1911 revision.

The American revisions of the Binet-Simon Scale include some or all of these tests. In the first American revision made by Goddard in 1911 (19), there are no significant changes in either the wording or placing of these three tests.

The Yerkes-Bridges-Hardwick point scale for measuring mental ability (73) includes only one of the type of test under discussion, i.e., the noting of differences between such objects as an apple and a banana, wood and glass, paper and cloth, for which a credit of six out of a total of one hundred possible points is given. The authors consider that the test involves the mental functions of analysis, comparison, and attention.

Terman's use of the tests of detecting differences and similarities as they appear in the Stanford Revision of the Binet-Simon Measuring Scale (51) is of interest, in that the revision was made with full consideration of practically all of the results which had been secured previously by workers in many countries, and in the light of results obtained by Terman and others specifically with a revision as the end in view.

He has retained each of the three original 1905 tests. Of test 16, "Comparison of known objects from memory," which is placed in the seven year group, he says, "Its excellence lies mainly in the fact that it throws light upon the character of the child's higher thought processes," since thinking means essentially "the association of ideas on the basis of differences or similarities." He finds

as Binet and Simon did that the ability to note differences precedes somewhat the ability to note resemblances. This test has undoubtedly established its value in a scale of general intelligence tests.

Terman has included test 20 of the 1905 series requiring the noting of similarities between two or more things, in a revised form. It will be remembered that Binet and Simon combined in the same test the giving of similarities between two, three, and four objects. Terman has found that the noting of similarities between two objects is satisfactorily accomplished by eight-year-old children, whereas the giving of similarities between three things is a sufficiently difficult test for twelve-year-old children.

The test requiring the giving of differences between abstract terms, which was a part of all three of the Binet-Simon scales is also retained in the Stanford Revision, as one of the "average adult" tests. Terman considers it a valuable one and denies that it is mainly a test of schooling. "The formation and use of abstract ideas, of one kind or another, represent, *par excellence*, the higher thought processes." He considers the main criticism that can be made of the test to be that "it imposes a somewhat difficult task upon the power of language expression."

To summarize, tests of the ability to note differences or similarities between objects not present to sense and differences between abstract terms, have played an important part in all forms of the Binet-Simon scales for measuring intelligence. Tests of this type in their simplest forms have been thought to be exercises in ideation, and to measure memory and the power of observation. It has been recognized that they make some demands upon vocabulary. This is especially true of the tests dealing with abstract terms. They have also been said to measure "natural good sense," "discriminative capacity to deal with knowledge already in possession of the subject." Other investigators credit them with involving the mental functions of analysis, comparison, and attention. Their chief value, however, is in their contribution to our knowledge of the associative processes and the light that they throw on "the character of the higher thought processes." All of these tests, in so far as they are dependent upon language, are of the "language comprehension" type. Their positions in scales

for measuring intelligence, which were determined experimentally, indicate their general relation to mental development.

Giving opposites in response to stimulus words. The opposites test in which the subject responds to a stimulus word with a word that is exactly opposite to it has had such a wide usage that it will not be possible to cite every study that has included it, but representative forms, attempts at standardization, and results of its application will be discussed.

Whether the stimulus words are "easy" or "hard," the opposites test is a test of controlled association. An attempt to analyze the mental capacities that it measures is complicated by the differences in response conditioned by the stimulus words. If the stimulus words are simple and thoroughly familiar, the response is given almost automatically and the test measures the rapidity of controlled association reaction time; if they are less familiar, or "hard," the test becomes in some degree, a measure of what Simpson (43) terms selective thinking, or what King and Gold (29) term "logical keenness in selecting the word which will express most nearly the contrary idea." They agree that a hard opposites test with young or immature or intellectually retarded subjects becomes a range of vocabulary test.

As early as 1905, Thorndike (54) made use of both easy and hard opposites tests in his measurement of mental and physical likenesses of twins. He found the tests to have a high degree of reliability.

List IV of Whipple's Manual (65) has been widely used by the earliest experimenters with opposites tests. Miss Norsworthy (36) has derived norms both for list IV and for its opposites for about 611 normal children of both sexes, between the ages of eight and sixteen and for an adult group. The medians of the number of correct associations for both forms of the list show a continuous increase with age, although the increase is more rapid from ages eight to twelve than from twelve to fifteen. The stimulus words are too easy to differentiate between the older individuals. The results of Norsworthy's experimentation with feeble-minded children (summarized by Whipple) show that no feeble-minded child reached the median performance of normal children; and only about one in a hundred was better than minus

1 P.E.; and only about six in a hundred were better than minus 2 P.E. of normal children of their age.

The efficacy of the opposites test in clearly differentiating between groups of varying intelligence has been demonstrated by other investigators and constitutes one of its chief claims to superiority as a test of "general mental ability."

The series of tests used by Bonser (6) in his study of the reasoning ability of children (385 boys and 372 girls) in the fourth, fifth, and sixth grades included three tests for controlled association, (*a*) sentence completion by supplying the word; (*b*) sentence completion through the choice of one of two words; (*c*) three sets of twenty words to which opposites were to be supplied (lists VI and VII and the opposites of list IV as given by Whipple). He says of these controlled association tests that the activity tested which is "vitally significant in all reasoning," is "that of spontaneity and accuracy in recognizing resemblances between the known of experience and the unknown of new situations;" and that the opposites test seems to be a test "of rather superior merit as a simple test for this general form of mental ability." He obtained a correlation of 0.85 between opposites and an average standing in all of the tests used to measure reasoning ability.

Squire (45), in her study of graded mental tests has included an opposites test and has selected the lists of stimulus words used by Bonser after canvassing the available tests of controlled association. Her subjects of which there were ten at each age from eight through thirteen, were pupils in the School of Education at the University of Chicago. Her method of scoring was rigid, and no partial credits were allowed. She recorded both time and response. Whipple has reproduced her table giving average time and number of responses for each age group, for each of the three lists separately and average together. The average number of responses for the three lists together increases rapidly from six to eight years and less rapidly from eight to ten. There is considerable increase at eleven, but no differentiation between eleven, twelve, and thirteen years. The average time of response follows somewhat the same order of improvement except that it shows a slight differentiation between the older group, that is, at eleven, twelve and thirteen years. On the basis of the results obtained

from these small groups of children, Squire has proposed norms for the opposites test.

Another early study employing opposites tests is that made by Simpson (43), in 1912, of two groups of adults representing as far as possible, "The two extremes of 'general intelligence' as judged by the world." The *good* group was made up of seventeen professors and graduate students at Columbia University; the *poor* group, of twenty men, all of mature ages, who had never held any position requiring a high grade of intelligence. Eleven of the twenty were staying at the Salvation Army Industrial Home, and seven were found in a mission on the Bowery. The remaining two were men earning comfortable livings, but who were recognized by their companions as being dull.

Simpson aimed to measure a variety of mental abilities which he grouped roughly under six headings, namely, "sense-discrimination, motor-control, efficiency in perception, efficiency in association, memory," and what for lack of a better term, he called "abstraction or selective thinking." His four tests of association were addition, associating words with hieroglyphic forms in pairs, adding letters to ma, be, ca, etc., to make words, and four lists of twenty easy opposites (three of which are found in Whipple's Manual, as lists IV, VI, VII, with some slight changes). In the opposites test, responses were given orally and scored for time in seconds, plus two seconds for a word half wrong, and four seconds for a word totally wrong or omitted.

As tests of selective thinking, he used the Ebbinghaus or mutilated text, the absurdities test, and four lists of twenty hard opposites. Responses were written. All of the words given as opposites were evaluated by three different judges on a scale from one to four. The final score given an individual was obtained by averaging the three evaluations. For each word totally wrong 36 seconds were added to the time score, and a proportionate number of seconds for each word partially wrong.

Simpson's results are interesting both as a corroboration of Norsworthy's finding that the opposites test differentiates clearly between groups of different intellectual status, and as they relate the opposites test to other tests, for both groups of subjects. The reliability of both the easy and hard opposites tests is satisfactory.

Simpson considers the easy opposites test a good measure of readiness of controlled association. He finds that with his *good* group, it is a measure of rapidity of association but in the case of a few of the *poor* subjects, it is very much like the hard opposites test for the more able subjects; that is, it seems to call for an exercise of their best thinking capacity and is thus a test of selective thinking. Also for the *poor* subjects the hard opposites test "becomes one of range of vocabulary rather than of ability to think up different opposites."

Both the easy and hard opposites tests separate the two groups completely. No member of the *poor* group surpasses the lowest score of the *good* group. No other test in Simpson's series makes as decisive a division as do these two tests, although the Ebbinghaus test is very nearly as efficient. Thus Simpson shows that the differences between these two groups are most strikingly revealed by language tests demanding selective thinking, and by language tests demanding speed and accuracy in easy association.

Both the raw Pearson coefficients of correlation and coefficients corrected for attenuation have been derived for all combinations of the tests used. On the basis of the raw coefficients, Simpson concludes that the grouping of the hard opposites and Ebbinghaus tests as tests of selective thinking is justified, but that the easy opposites test is more nearly related to the tests of selective thinking than to the other association tests with which it was originally grouped.

This study of Simpson's is a careful and painstaking piece of work. It is obvious that its value would have been multiplied many times had he employed a larger number of cases in both groups. But despite the limited number of cases, very definite trends have been established. Opposites tests do differentiate between groups of different intellectual status; they do demand of a subject an ability to make fine discriminations and to judge between possible responses, hence to think selectively; they correlate highly with other language tests; the individual stimulus words must be taken into account whether they are easy and commonplace, or less familiar and difficult.

The results of other investigators who have used opposites tests will be summarized more briefly. Whipple has given a

summary of studies appearing through 1915. Pyle (41) has furnished norms for year groups from eight through eighteen years and an adult group, for the number of opposites written in 60 seconds in a group test, using list IV (the most commonly used list of easy opposites). The increase in score is more regular for girls than for boys, and is somewhat higher for girls than for boys at every age except at eight and nine. The 1920 revision of Pyle's Manual (42) furnishes norms for the average number of correct responses to two lists of 50 opposites each. The norms are given separately for the two lists, for city and country children, and for boys and girls. For the first list, the age range is from eight through twelve years; for the second list, it is from ten through eighteen years. The girls exceed the boys at practically every age and the city children surpass the country children at all ages.

Carpenter (13) has derived norms for time and errors for list VII, given individually to age groups ranging from seven through fourteen and varying in size from 7 to 58 cases. He states that because of irregularity in the giving of the test, his norms are not reliable below age nine. From ages nine through fourteen, there is a regular decrease in average time of response and a fairly regular decrease in average errors.

Woolley and Fischer (71), in their study of over 800 children fourteen years of age, which included an opposites test of 20 words to which the responses were written, found a marked and consistent positive correlation between this test and school grades. The superiority of the public school children over the parochial school group was most marked in the opposites and puzzle box tests, which tests were considered by Woolley and Fischer to be least influenced by school drill.

Kitson (32) in a study of 80 college students, in which sixteen tests were used, derived rank difference coefficients of correlation for each of the single tests with the net score. The two equivalent lists of opposites of Woodworth and Wells which were used yielded a coefficient of 0.53 ± 0.08 with the net score which was the second highest of any of the correlations between a single test and the net score.

King and M'Crory (30) tested 276 women and 268 men, fresh-

men at the University of Iowa, with a series of seven tests including completion, arithmetic, analogies, information, visual imagery, logical memory and opposites (both easy and hard). Their lists of opposites were based on Simpson's four lists of hard opposites, rearranged into two lists of forty words each, in order from the easiest to the most difficult. They derived the coefficients of correlation between each test with each other test, with the test average and with grades, for men and women separately.

The opposites test correlated most highly of any of the tests with university grades for both men and women. The coefficients of correlation of the opposites tests with the other measures are as follows:

	276 WOMEN	268 MEN
Completion (Simpson)	0.31	0.79
Arithmetic speed (Courtis)	0.03	Neg.
Arithmetic accuracy (Courtis)	0.01	Neg.
Analogies (Whipple)	0.52	0.77
Information (Whipple)	0.24	0.56
Visual imagery (painted cubes)	0.07	0.56
Logical imagery (Whipple)	0.32	0.38
Test average	0.51	0.88
University grades	0.45	0.84

The correlation coefficients for the men are consistently higher than those for the women. We should have to know something about the distributions for the two groups before attaching any particular significance to this difference. The coefficient of 0.84 between opposites and university grades is unusually high for any test with grades. King and M'Crory conclude that if they were to choose any one of the tests as a sign of the mental ability of the individual being tested it would be the combination of the two opposites tests which they used.

In a study of a group of 100 freshmen at Northwestern University, Uhl (60) administered three tests each of which he correlated with each other and with first semester grades in English and mathematics. His three tests were: (*a*) Trabue completion,

scales K and M; (*b*) hard opposites, a list of twenty words, neither given nor described; (*c*) an information test based on Whipple's list of one hundred words. The coefficients of correlation between opposites and other tests and grades are as follows:

OPPOSITES WITH	*r*
Trabue Scale K	0.21
Trabue Scale M	0.26
Information	0.26
English grades	0.44
Mathematics grades	0.29

These coefficients are lower than those usually found. It is possible that the opposites test was too simple to differentiate between as highly selected a group as college freshmen.

Baum, Litchfield and Washburn (1) report a coefficient of 0.30 ± 0.11 for the Woodworth and Wells hard opposites test with academic rank, in a group of 38 seniors, chosen to form a fairly continuous scale from the highest to the lowest in the class. This coefficient is not of particular significance in the light of the selection of the group. Of more interest is their comparison of the two groups of 25 students each, all fifty of whom were up to the standard required for graduation, but who represented the highest and lowest ends of the total distributions of college seniors. "While neither the opposites test nor the analogies serves entirely to separate such groups, there is a distinct correlation in the case of both between test performance and academic record."

In a study of 200 college freshmen, Carothers (12) employed a series of nineteen tests, including both motor and language tests, given individually. A combination of the Woodworth and Wells lists of opposites was used. Each subject was required to give the correct opposite before going to the next stimulus word. If a wrong response was made, therefore, she was stopped and required to give the correct one. The score was the total time for completing the task. This test was included in the series as a measure of "facility in handling words," and as a "test which would indicate 'general tendency' or 'adjustment to react according to instructions' and also (to) measure the quickness and accuracy

of association of ideas." On the basis of the inter-test correlations and correlations with other tests in the series, certain tests group together. As could have been predicted, the opposites test is related most closely to other verbal tests, tests of selective judgment, such as the completion, mixed-relations, and word-building tests. The third highest correlation between two tests was 0.57 between opposites and mixed-relations. Carothers reports the following correlations between opposites and grades in college subjects:

OPPOSITES WITH	CASES	r
Language	97	0.17
Mathematics	88	0.01
Science	41	0.33
Philosophy	27	0.01
History	26	0.30

The coefficients are small as are most of those reported in this study. The Woodworth and Wells lists of words are obviously too simple to involve any "selective thinking" and association time, alone, does not seem to correlate to any considerable degree with school subjects.

Tolman (56) has reported an interesting and suggestive study of the relation of certain tests to students' estimates of their own English and mathematical ability, although his groups are too small to justify completely his use of the correlation method. Seventy members of an introductory psychology class were asked to tell whether they thought they had more ability in English or mathematics. They were then given a series of eight group tests including opposites (King and Gold), Trabue completion, rhymes, and disarranged sentences which were supposed to be "English tests" and arithmetic (mental problems), algebra, generalizations and problems to be solved mentally, which were "mathematics tests." Tolman has omitted mention of the opposites and generalization tests in his report of the tests which he considers "most successful" presumably because of their low correlation with other tests. Less time was allowed for opposites than for any other test which may explain in part the low degree of correlation.

A summary of the attempts to standardize opposites tests, up to 1921, is given by Means (35).

One of the earliest, if not the earliest, attempts to standardize lists of stimulus words suitable for opposites tests was made by Woodworth and Wells in 1911 (70), who as a result of experimentation made up two lists of twenty words, supposed to be of equal difficulty, and of which the two halves are equivalent. A third list was made up of the easiest words in the first two lists. The words were selected in order that the subjects might have a perfect score, and the test is scored on the basis of time. Some of the uses of the Woodworth and Wells lists have already been mentioned.

A further attempt at standardization was made by King and Gold in 1916 (29). Simpson's eight lists, totaling 160 words, were presented individually to 9 faculty members, 23 graduate students, 47 seniors, and 21 juniors of the department of education and psychology in the University of Iowa. The time was recorded for each list separately and the responses taken down in shorthand. On the basis of the hundred records for each stimulus word, the per cent of failure was computed. For each stimulus word were recorded: (*a*) a value in terms of percentage of accuracy; (*b*) the acceptable responses; (*c*) the frequency of each response. On the basis of these data two lists were assembled, a list of 40 easy opposites, the responses to which were 90 per cent correct; and a list of 25 hard opposites, the responses to which were less than 69 per cent correct. Two additional tests were given and some of the inter-test correlations determined. The coefficients obtained were as follows:

Hard opposites with information (Whipple)	0.40
Hard opposites with vocabulary (Whipple)	0.19
Easy opposites with information (Whipple)	0.37
Easy opposites with vocabulary (Whipple)	0.16
Information with vocabulary	0.45

The first attempt to assign to each stimulus word a point value based on its relative difficulty was made by Greene in 1918 (20). He represented one half of the eighty words used by King and Gold to 990 freshmen, and the remaining half to 710 freshmen. The responses were given values of 0, $\frac{1}{2}$ and 1. The percentage of

failures for each word was then determined by allowing a value of one unit for each correct response, and one half for a half correct response and subtracting the total from the total possible right (990 and 710). To quote from Means's summary,

> By reading directly from the table based upon the area of the probability curve and assuming that the base line is broken arbitrarily at 3 sigma, the percentage scores were changed into percentile values. These values were then totaled, and each value in turn divided by the total, thus converting the percentile values into relative point values. The points were based on accuracy alone, no account being taken of the time.

The standardization made by Means (35) takes account of both accuracy and time of response. Three hundred twenty-three words used by previous experimenters and 40 original ones were divided into two lists of about 150 words each. Commonly misunderstood words, words with an accuracy score of 100 per cent, and words which had an opposite formed by adding the prefix "un" in frequent and reputable use, were weeded out. The words were given orally, and the reaction times recorded in fifths of a second with a stop watch, together with the response. The final list consisted of 68 words. The responses which were given by over one hundred college students were scored 1, $\frac{1}{2}$, and 0, according to the average judgment of five judges. To arrive at the relative difficulty of the stimuli, the percentage of failures was multiplied by the median time. Thus 68 words were arranged in the order of difficulty and a value ranging from 1 through 26 assigned each word.

The list in its final form was given as a group test, with a time allowance of six minutes, to 1628 college students. The results show a decided increase in score at each percentile from the freshman to the junior year, but there is less separation between juniors and seniors, which might be expected since a college class is fairly stable by the time the junior year is reached. There is a marked differentiation between seniors and graduate students. From her data, Means feels justified in concluding that the opposites test as here given, at least, is one in which success depends more upon native ability than upon number of years schooling.

There remains to be mentioned but one other attempt at standardization, made by Van Wagenen in 1920 (61). Approximately

three hundred stimulus words were given orally in sets of fifty each, at the rate of four seconds for each word, to 148 sixth and seventh grade children and 54 college juniors and seniors. All of the responses were submitted to thirteen advanced students to rate. The frequencies of the sixth and seventh grade responses were weighted by three in the case of the more difficult stimuli, and two in the case of those of medium difficulty. The frequencies for the college students were similarly weighted. The weighted difficulties of the two groups were then combined, a value assigned to each word, and four equivalent lists made up on this basis. The intercorrelations between the four lists range from 0.68 to 0.76.

To summarize, the opposites test in which a word must be given in response to a stimulus word, opposite in meaning to the stimulus word, has had a long and varied history. The uses to which it has been put and the results of its application are conditioned by the nature of the stimulus words, in relation to the age or development of the subjects tested. If the stimulus words are simple so that opposites can be supplied to all of them, the test becomes one of controlled reaction time and measures spontaneity and accuracy of recall, and the efficacy of old associations. The application of this type of test to young children, for whom the stimulus words are naturally more difficult than for adults, has shown rather definite age differences in performance (Squire, Pyle, Carpenter, Woolley and Fischer). The efficacy of the test in separating groups of inferior intelligence from groups of superior intelligence was shown by Norsworthy and Simpson. It has been found to be less influenced by school drill than many other tests.

The results of applying the test to more advanced subjects, as college students, show that the test, if merely one of controlled association reaction time, does not correlate highly with other verbal tests nor with such a criterion as grades (Carothers, Uhl).

If the stimulus words are difficult, so that the nature of the response is as important as the time of response, the test becomes more nearly a language test, or in some instances, a measure of range of vocabulary. It has also been considered as a test of "mental efficiency," of "selective thinking," of "logical keenness." It has been found to have a high degree of reliability, to differ-

entiate between groups of varying intelligence, to correlate more highly than information, arithmetic, or completion tests with grades.

Thus far, this summary has considered opposites tests in which the difficulty of the individual stimulus words has not been determined. Attempts to standardize individual stimulus words, both as to nature and speed of response have been made. As yet, results of application of the test in this form are not numerous. One experimenter found an increase in efficiency measured by the test with an increase in years in college which she attributed not to training but to the weeding out of inferior material.

Synonym-antonym test. When psychologists were mustered in 1917 to devise tests for use in army mental testing (74), the field of tests was canvassed, with twelve criteria in view. Among them, in addition to certain criteria of economy in time and energy of scoring, were validity as a measure of intelligence, range of intelligence measured, unfavorableness to coaching, malingering, and cheating, and independence of schooling. Thirteen tests, including the synonym-antonym, opposites, and vocabulary tests, were proposed as likely material.

These original thirteen tests were carefully judged and rated by five psychologists in relation to the twelve criteria and reduced to ten tests. The synonym-antonym test was retained as test six in the series, whereas the vocabulary and opposites tests were both eliminated on the basis that "the vocabulary test, the synonym-antonym test, and the opposites tests were of the same general type. The synonym-antonym test embodies most of the advantages of both the vocabulary and the opposites test and it has the advantage of requiring less time and being easily scored."

The task of preparing the items for ten equivalent lists of forty pairs of words for the synonym-antonym test was assigned to Terman. The words used in the lists were all taken from a "vest pocket" dictionary (Funk and Wagnall's) "in order to guard against the inclusion of rare or technical terms." Out of an original five hundred pairs, some were arbitrarily eliminated by vote of the committee as being unfit for use. From the remainder, forty pairs were drawn by lot for each of the ten lists.

In the work of revising the series of ten tests in Alpha, the

synonym-antonym test was found to yield too many zero scores, but was otherwise "one of the best tests in the scale." A histogram showing the distribution of scores for an "experimental" group made up of 1047 men from eleven camps is found on page 625 of volume XV of the *Memoirs of the National Academy of Sciences*. The most striking feature of the histogram is the piling up of scores at the lower end, at 0 and 1. The three causes of this skewed distribution as given by Brigham (9) are, "first, the illiterate group could not attempt it; second, the stupid and literate could not understand the instructions and could not make the kind of judgment demanded." The third cause is connected with the method of scoring, which is such that in the long run chance responses would give scores around zero.

For the army group at least, the synonym-antonym test differentiated very well between high and low grade individuals. It was considered one of the most effective tests in differentiating officers from enlisted men, and enlisted men from the feeble-minded. The efficacy of this test in differentiating between groups of different intellectual status suggests similar results of Norsworthy (36) and Simpson (43) with the simple opposites test and, in a less direct way, those of Chapman and Dale (14) with the synonym-antonym test, which will be discussed in some detail later.

Since its use in the army tests, the synonym-antonym test has been included frequently in batteries of group tests. An examination of the sub-tests of 29 group examinations (which does not claim to exhaust the list of group tests), for use in grades ranging from one through fourteen and the first year of college, reveals the interesting fact that 22 of the 29 series employ opposites tests of some sort and of the 22, 10 are of the "synonym-antonym type.[1] A further examination of the stimulus words which make up these tests yields the additional fact that the army tests have been freely drawn from which is not surprising, however, since Terman put into his ten lists the best of the available pairs of words.

The value of this test has been demonstrated in the army

[1] Most of these group tests are listed by Whipple in the 21st Year Book of the National Society for the Study of Education (59).

testing, and it is assumed that the devisors of batteries of tests have further reassured themselves of its value in making their selection of tests, but published experimental data are very meager. There are, to the writer's knowledge no published distributions with the exception of that for the army data although it is to be expected that groups differing from the army group may show interesting differences in distribution. Practically the only data are in the form of correlations of sub-tests in group examinations, which in turn, are not very numerous. The few published studies will be summarized.

Jordan (25) (26) has published two studies showing the interrelations of four group tests, and their elements, and their relation to grades in school subjects, as indicated by coefficients of correlation. In both studies the subjects are high school pupils, 67 in the first study and 64 in the second, presumably the same individuals. The four intelligence tests used with all of the subjects were Army Alpha, Otis Group Intelligence Scale, Terman Group Test of Mental Ability, and Miller Mental Ability. The sub-tests of each of these series are listed below together with the working time allowed. The names of the tests merely indicate their general nature:

Army Alpha

	minutes
1. Oral directions	2½
2. Arithmetic	5
3. Reasons	1½
4. Synonym-antonym	1½
5. Mixed sentences	2
6. Number completion	3
7. Analogies	3
8. Information	4

Otis

1. Following directions	5
2. Opposites	1½
3. Disarranged sentences	1½
4. Proverbs	6
5. Arithmetic	6
6. Geometric forms	6
7. Analogies	3

8. Similarities 4
9. Completion 6
10. Memory 3

Terman

1. Information 2
2. Best answers 2
3. Synonym-antonym 2
4. Logical selection 3
5. Arithmetic 4
6. Sentence meaning 2
7. Analogies 2
8. Disarranged sentences 3
9. Classification 3
10. Number completion 4

Miller

1. Mixed sentences 8
2. Cause and effect 5
3. Analogies 6

Among the 31 sub-tests there are two synonym-antonym tests, Alpha 4 and Terman 3, and one opposites test of a different form, Otis 2.

The coefficients of correlation of the four tests and their 31 sub-tests with grades in English, general science, and history, and average grades are published in full in the 1922 report and may be summarized with special reference to the opposites tests as follows: No one of the opposites tests is included among the five best single tests (judged purely by size of correlation coefficient) out of a total of 31 tests, for predicting success in average high school grades for an entire year, or in mathematics, general science, or history grades, although they are less predictive of success in mathematics. The two synonym-antonym tests stand second and fifth among the five best tests for predicting success in English. The groups on which these conclusions are based are very small, but tendencies are at least indicated. The classification of opposites tests as "language" tests would seem to have some justification in these results.

Jordan's second report (26) which is of greater interest, relates the same four tests and their 31 sub-tests to a series of criteria, *i.e.*, Stanford-Binet mental age, teachers' estimates of intelligence

(an average of four estimates), a composite score of the four tests, a learning test devised by the author, and chronological age. The coefficients of correlation for the three opposites tests, Alpha 4, Otis 2, and Terman 3, followed by the five largest coefficients for each criterion will be reproduced, together with their rank order position among the total 31 sub-tests.

The correlation coefficients with the Stanford-Binet mental age are as follows:

1. Alpha 4, synonym-antonym	0.600
2. Otis 2, opposites	0.577
3. Terman 3, synonym-antonym	0.553

1. Alpha 4, synonym-antonym	0.600
2. Otis 2, opposites	0.577
3. Terman 3, synonym-antonym	0.553
4. Otis 4, proverbs	0.552
5. Otis 5, arithmetic	0.541

The three opposites tests stand above all other sub-tests in degree of correlation with mental age, as measured by the Stanford-Binet tests. Inasmuch as it has been very frequently asserted that the Stanford-Binet mental age is the best available criterion of general intelligence, the claims that have been made for the opposites test as a good test of intelligence are validated to some extent. On the other hand, the relationship between the two may be due to a common verbal element.

The coefficients of correlation with teachers' estimates are as follows:

7. Alpha 4, synonym-antonym	0.553
16. Otis 2, opposites	0.459
3. Terman 3, synonym-antonym	0.584

1. Alpha 2, arithmetic	0.622
2. Miller 2, cause and effect	0.611
3. Terman 3, synonym-antonym	0.584
4. Otis 7, analogies	0.567
5. Otis 8, similarities	0.559

The two synonym-antonym tests again correlate more highly with the criterion than the Otis opposites, but the opposites tests as a whole correlate less well than some of the other tests.

The coefficients of correlation with a composite of the four tests are as follows:

4. Alpha 4, synonym-antonym	0.763
7. Otis 2, opposites	0.755
1. Terman 3, synonym-antonym	0.809

1. Terman 3, synonym-antonym	0.809
2. Miller 2, cause and effect	0.799
3. Alpha 8, information	0.771
4. Alpha 4, synonym-antonym	0.763
5. Otis 4, proverbs	0.755

Although it is impossible to interpret accurately coefficients of correlation of sub-tests with composite scores, without a careful analysis of the weight of each test in the composite, in order that the spurious correlation can be allowed for, still one is tempted to attach some significance to a coefficient of 0.809 (Terman 3) as contrasted with one of 0.409 (Otis 3). In this instance, the fact that the Terman synonym-antonym test correlates most highly with the composite score and the Alpha synonym-antonym fourth highest, may simply indicate that the composite is heavily weighted with language tests, or tests involving the understanding of the meanings of words. Some light would have been shed on this question had Jordan derived inter-test correlation coefficients.

If the rank order positions of the five highest correlations with each of the criteria are averaged, the opposites test stands out clearly ahead of the rest of the tests, representing seventeen types of test material. Arithmetic problems come second, followed by geometric figures and proverbs.

All of the available correlational data up to 1923 concerning each of the four group tests have been summarized by Jordan. His summary tells nothing of the size or homogeneity of the groups represented so that little of any value can be deduced from it. There is an indication that the synonym-antonym test, especially as in Alpha, correlates more highly with mental age and average grades than any other test in the Alpha series. The summary serves, however, to point out the meagerness of the available data on sub-tests, and to suggest the need for a careful analysis

of such tests, in relation to the nature and distribution of the cases in the groups studied.

Further justification for the inclusion of the opposites test as a language test is found in a study made by Kelley (27) of data obtained from 1257 Virginia elementary school children on the original try-out battery of the National Intelligence test and of the Stanford achievement tests, which included such tests as arithmetical reasoning, information, synonym-antonym, sentence completion, vocabulary, practical judgments, computation, etc. Kelley concludes on the basis of correlations that such tests as written directions, opposites, practical judgments, vocabulary, and completion, do not test disparate functions, in spite of the fact that they are differently labeled, but are really "all tests of the same thing." In the case of elementary school children they do not measure five capacities but one. He designates them as verbal tests and compares them with arithmetic tests, and finds that distinctions can be drawn between arithmetical computation ability and general power of logical analysis and reasoning with verbal material; between arithmetical computation ability and knowledge of words; between arithmetical reasoning ability and knowledge of words.

Chapman and Dale (14) selected from 5000 National Intelligence Test blanks (Scale A), two groups of blanks representing children of British or American birth. The first group was made up of children under ten years old (designated the Young Bright or Y.B. group); the second group was made up of children thirteen years of age or over (designated the Old Dull, or O.D. group). The scores for the members of both groups fell between 70 and 119. For each total score in the Y. B. group there was an equal score in the O.D. group. Fifty pairs of papers were obtained. A comparison of the performance of the two groups in the different tests in the series was then made. The performance of the O.D. group exceeded that of the Y.B. group in arithmetical reasoning, symbol-digit, and logical sequence; the two groups were practically equal in the sentence completion test. The most marked difference in performance was in the synonym-antonym test, in which the O.D. group average 12.6 and the Y.B. group, 19.0. If the relation of the performance of the Y.B. to the O.D. group is ex-

pressed as a ratio, Y.B./O.D., the results will be clearer. The ratios for each test and the total score are as follows:

1. Arithmetical reasoning	0.87
2. Sentence completion	1.02
3. Logical sequence	0.92
4. Synonym-antonym	1.51
5. Symbol-digit	0.85
Total Score	1.00

These results indicate that success in the synonym-antonym test depends to a higher degree on native intelligence than on experience, since the bright children are at least three years younger than the dull children.

Thus, the synonym-antonym test has been found to differentiate between different grades of intelligence more clearly than these other tests. Two experiments of a different nature have indicated that success in the synonym-antonym test is more dependent upon native ability than upon experience. For high school pupils, the synonym-antonym test has been shown to correlate more highly with English grades than with average grades, which indicates a specialization of the test as a language test. In comparison with seventeen types of test material, the synonym-antonym test was found to correlate most highly on an average with such criteria as Stanford-Binet mental age, teachers' estimates of intelligence, and a composite of intelligence tests. The synonym-antonym test in the Alpha series, as applied to college and university students, correlates more highly with grades than any other sub-test in the series.

Summary of the three types of tests for the measuring of ability to distinguish differences and similarities. Tests of the ability to note differences or similarities whether they are of a form in which differences or similarities between objects or abstract concepts have to be noted, or in which an opposite has to be supplied, or in which a relationship between two words has to be judged either similar or opposite, have all tended to differentiate between groups of different mental development. It is obvious, then, that the ability to note differences or similarities is a fundamental factor in what is designated "general intelligence." Depending upon the exact nature of the test material, these tests have been variously thought

of as "exercises in ideation;" as measures of "natural good sense;" of "discriminative capacity to deal with knowledge already in possession of the subject;" as involving the mental functions of analysis, comparison and attention; as throwing light upon the character of the higher thought processes; as tests of controlled association and as measures of spontaneity and accuracy in recall and of the efficacy of old associations; as tests of selective judgment or reasoning, of logical keenness; as language tests measuring the range of vocabulary.

With groups of elementary and high school students this type of test has been considered as but slightly influenced by school drill or amount of experience, and less so than many other tests. If the stimulus words are of sufficient difficulty to offer a real problem to the subject, it has been found to correlate relatively well with average grades, but better with English grades and considerably less well with mathematics grades.

In so far as this type of test can be designated a language test, it would seem to be a measure of the ability to understand rather than to compose or construct language.

b. Completion tests

The completion test, at least in one of its forms, has probably been referred to specifically as a test of "language ability" more often than any other single test. Ebbinghaus (18), the originator of the completion test method, presented to his subjects a prose passage, mutilated by the elision of letters, syllables, words and phrases, and required them to restore the passage. He termed this the "Combinations Methode" and has characterized it as a "real test of intelligence," which he says, "consists in the elaboration of the whole into its worth and meaning by means of many-sided combination, correction, and completion of numerous kindred associations," and as a "simple, easily applied device for testing those intellectual activities that are fundamentally important and significant both in the school and in life" (quoted from Whipple).

Whipple (65) has recognized the difficulty of classifying this method in a system of tests, because of the dependence of the

mental processes tested upon the number and kind of elisions made in the text. He says,

> To take extreme cases, if the elisions are numerous and sweeping, it may become a linguistic puzzle of a very difficult variety, and it then belongs rather in the group of tests of active or creative imagination of the literary type; if, on the other hand, the elisions are but few and simple, it may degenerate into a simple test of controlled association of any desired degree of ease. Again, if the original text be first read to the examinee, as some suggest, the test becomes in the main a test of associative recall, *i.e.*, a form of memory test.

Terman (47) in his study of genius and stupidity in which he undertook to measure the abilities of seven bright and seven dull children at eight more or less different points, included "mastery of language" as an important point for the reason that "language growth epitomizes the development of a child's intelligence as a whole." His observations in this field included reading, building words from given letters, correction or completion of mutilated text, spelling, fluency of expression, and facility in obeying oral commands.

Two somewhat different tests patterned on the Ebbinghaus method were given, one in which letters, syllables, and words were elided, and a second one which was first read to the subject, and from which words alone were elided. With certain exceptions the two groups of subjects were widely separated in each of the two tests. Terman's experience with this test causes him to regard it favorably. It does indicate something as to the command of language, but he is inclined to think that "somewhat mechanical activities like memory or association, as distinguished from synthetic or combinative processes, play a relatively more important rôle in this test than Ebbinghaus assigns to them." He also thinks that success in the test depends on the degree of acquaintance with the sort of literature presented in the text, and perhaps still more "upon peculiarities of language development in the subject."

Terman and Childs (53) have made a tentative standardization of the completion test, of the same type and given norms of performance for ages nine through fourteen. Their results show clear differences in performance from year to year, except in the twelfth year, which showed slight gain over the eleventh year. They conclude, therefore, that the test fulfills the most important

requirement for use in a measuring scale of mentality, and offer as their opinion that it brings to light fundamental differences in the thought processes.

Ebbinghaus held that his test correlated well with intelligence. Similar results were obtained by Cohn and Dieffenbacher (15), and Wiersma (68), although they did not employ the correlation method to express the degree of relationship. Correlation coefficients have been reported for the Ebbinghaus test in the form under discussion by Brown (19) who found a coefficient of 0.43 with one group of 66 boys aged eleven to twelve years, and 0.69 with another group of 39 girls, aged eleven to twelve; by Burt (11), who found coefficients of 0.48 and 0.53 with two tests and intelligence, with a group of about 60 to 75 boys and girls between the ages of 11.5 and 13.5; by Wyatt (72), who found a coefficient with a subjective estimate of intelligence of 0.85 for one group of 34 boys and girls, aged eleven to thirteen and 0.61 for another group of 41 boys and girls, aged ten to twelve. Simpson (43) found that this test differentiated very definitely between a group of subjects of superior intelligence and another group of distinctly inferior intelligence.

These coefficients are sufficient to indicate that there is a relatively high degree of correlation between this test and intelligence. It is doubtful whether the exact size of the coefficients can be depended upon since earlier studies employing the correlation method have not, as a rule, taken into account the heterogeneity of the group tested, nor has the age factor been entirely eliminated.

In a study of more than 750 college freshmen, Bell (3) gave the following nine tests, the net working time for which was twenty minutes: 1, Cancellation of triangles (Simpson); 2, Addition (Simpson); 3, Association (Thorndike's learning pairs); 4, Recognition (Simpson); 5, Selective judgment (Bonser); 6, Easy directions (Woodworth and Wells); 7, Hard directions (Woodworth and Wells); 8, Alternatives (Squire); 9, Completion (Terman and Childs). Inter-test correlations and correlations with grades were derived. Although the coefficients are all very low, the highest degree of correlation between any test and grades is that of completion with English, which is 0.31. The

highest inter-test correlations exist among the two directions tests, alternatives, and completion, probably due as Bell suggests to the linguistic factor which is present in all of them.

If the coefficients of correlation of each test with each of the remaining tests are averaged separately for each test, the completion test shows the highest average, which indicates that it has the most in common with the largest number of tests and is probably the best of the tests represented in this study as a measure of a general ability of some sort.

Binet and Simon (5) have included in their 1905 scale a test, "Verbal gaps to be filled," patterned after the Ebbinghaus test, and intended as an "exercise of judgment." In their subsequent scales, they have substituted a test which was inspired by the investigations of Ebbinghaus, and which consists in rearranging a group of words to make a sentence. It is placed in the eleven year group in the 1908 scale, and in the twelve year group in the 1911 revision. They have offered no new suggestions as to what the test in this form is to measure.

The disarranged sentences test, as it is generally designated, has been frequently used in series of tests for measuring intelligence. It has been retained in the Stanford Revision, in the twelve year group. Terman (51) thinks that success in this test depends upon "the ability of intelligence to utilize hints, or clues" which in turn, "depends upon the logical integrity of the associative processes."

Goddard (19) and Kuhlmann (33) have also retained this test in the eleven year group as it was placed by Binet and Simon in the 1908 scale.

The most extensive single contribution to the development of the completion test has been made by Trabue (58). He was interested in deriving one or more scales for the measurement of ability along certain lines closely related to language. As he recognized the difficulty if not the impossibility of standardizing paragraphs and of finding two paragraphs of equal difficulty, he devised scales, made up of sentences graded in difficulty. After extensive experimentation first with the "Graded Series," made up of 56 sentences, and then with Scale A, composed of twelve pairs of sentences, he finally arrived at a series of short scales,

composed of from seven to ten sentences each, the scoring of which has been standardized.

Very few data pertaining to the relation of the completion test to other tests are presented in Trabue's monograph. Of the coefficients of correlation given, only those will be referred to which have taken into consideration the age factor. For 30 seventh grade children, the coefficients of correlation for Language Scale A and other measures were as follows:

With English Composition (Hillegas Scale)	0.72
With Courtis Problems in division	0.04
With Thorndike's reading scale Alpha	0.47
With Thorndike's reading scale A	0.49

For 33 sixth grade pupils, the coefficients were:

With teacher's estimate of intelligence	0.74
With Woody multiplication test	0.51
With Courtis multiplication test	−0.12
With Thorndike's reading scale Alpha	0.47

For 29 sixth grade pupils the coefficients were:

With a combination of all marks made in one semester	0.49
With a combination of Thorndike's scale Alpha, Buckingham's spelling scale, and Hillegas composition scale	0.58
With a combination of vocabulary, opposites, mixed relations, and proverbs tests	0.39

The groups represented are too small to yield conclusive results. The coefficients suggest that there is a positive relation between ability to complete sentences and ability in other tests of language and general intelligence.

Some of the uses to which the Trabue Completion Scales have been put and the results of their application will be summarized.

Pintner and Paterson (39), in a study of the language ability of deaf children, selected the Trabue Language Completion test because of its combination of the two factors of comprehension and composition, which made it, in their opinion, the best single test for the "best all round measurement of language ability."

Whipple, Henry, Manuel and Coy (67) in their study of gifted children applied a series of 55 tests to two groups of children, a

selected group of superior fifth and sixth grade pupils and a control group of fifth and sixth grade pupils. Twenty-seven of the tests were studied with special reference to their efficiency in separating the special from the control group. Nine of the tests were found to be especially efficient, among them the Trabue Completion Scales B and C, J and K.

McCall (34) reports the results of the examination of 88 sixth grade pupils with a number of tests including Thorndike's Visual Vocabulary (30 minutes); Thorndike's Reading Scale Alpha (30 minutes); Trabue Completion (30 minutes); Arithmetic (six specially selected problems, 30 minutes); Omnibus I, including easy and hard opposites, verb-object, supra-ordinate, mixed relations, easy and hard directions, and addition (30 minutes); Omnibus II, which tested reasoning ability, ability to give opposites to certain hard words, to give verbs to specified subjects, to add proper letters to complete unfinished words (30 minutes); Proverbs (Ruger); and a series of practice tests. He also obtained such measures as age, school marks and teachers' estimates of intelligence. A composite score was derived which combined according to certain weights all of the measures described with the exception of proverbs and one of the practice tests. According to coefficients corrected for attenuation, the two Omnibus tests combined, and completion show a practically perfect correlation with the composite. The composite score was considered as the best measure of mental ability and the value of each of the other measures was related to it. Thus McCall lists the seven "best measures of mental ability," together with their corrected coefficients of correlation with the composite as follows:

Omnibus	1.00
Completion	0.96
School marks	0.91
Teachers' estimate	0.86
Reading	0.81
Visual vocabulary	0.80
Arithmetic	0.72

It is of interest that one specific test, the completion test, correlated almost as highly as a group of tests, represented by the

omnibus test, with a combination of as varied measures as go to make up the composite.

Pintner (38) used the Trabue Scales B and C to obtain a measure of the increase in language ability of school children. His subjects were 598 children in a junior high school, grades 2B to 9A; and 418 in a grade school, grades 2B to 8A. Scale B was given in October and Scale C in the following May. He found that the difference in language ability and in language progress of the two schools was marked and constant, which is of interest since the junior high school, which surpassed the grade school in ability and progress, was considered a rather superior school, whereas the grade school would be ranked as an average city school. The superiority of the junior high school in language ability as measured by the Trabue scales may be due either to superior environmental conditions or to superior intelligence.

Colvin (16) reports the results of psychological tests at Brown University, administered to entering students. Two equivalent tests, Brown Series I and II, composed of the following sub-tests were given: Test A, ten mutilated sentences taken from Kelley's standardization of the Trabue test; Test B, twenty-five of the difficult words from Terman's vocabulary test; Test C, analogies; Test D, reasoning. Forty-five minutes were allowed. The individual tests were weighted in the total score so that Test A was assigned the most importance and Test D the least. Two hundred twelve men took Series I, 178 of these took Series II as well, and 103, Army Alpha. Some of the coefficients of correlation derived by Colvin are as follows:

Average of Series I and II with first term grades	0.53
Test A with first term grades	0.41
Test B with first term grades	0.45
Test C with first term grades	0.52
Alpha with first term grades	0.43
Alpha with Series I and II	0.51

The completion test correlated about as well with the first semester grades as did Alpha, but less well than analogies and vocabulary tests. The fact that the individual tests are as predictive of academic success, at least for this group, as the entire Alpha, is interesting evidence of the comparatively close

relationship between predominately language tests and academic success.

The results of Uhl (60) indicate that the completion test is more predictive of success in English than in mathematics, whereas those of Tolman (56) indicate that it is equally predictive of grades in both subjects. As Tolman has based his conclusion on only 29 cases, but little weight can be attached to it.

An interesting study demonstrating the immediate relation of the Trabue completion test to definite language training was made by Henmon (21). During the first semester, a sophomore high school class of 54 was given an intensive course in word study and analysis as a substitute for the regular work in English composition and literature. Tests were then chosen to measure increase in ability along the lines tested as follows: 1, Terman's vocabulary test of 100 words to measure increase in vocabulary; 2, Thorndike's visual vocabulary test to measure increase in ability to give meanings accurately; 3, a special test of twenty-five words where a knowledge of roots would help; 4, Trabue Completion Scale L to measure ability to discriminate in the choice of words; 5, test Ia and Ib of Thorndike's Intelligence Examination, Part III, to measure the ability to read a difficult prose passage understandingly. These tests were given to over 350 sophomore students. A group was then selected from the untrained students to pair with each member of the special group, equal in scholarship, based on freshmen marks and marks of the first semester of the sophomore year, and of equal amounts of language work in French and Latin. Whereas, in average grades the non-word study group was slightly superior to the word study group, the latter excelled the former in all of the tests by a significant difference. In four of the tests, Thorndike reading, Trabue completion, word meaning, and Terman vocabulary, the difference between the two groups was four times the probable error of the difference. That the special group excelled in the Trabue test, is more significant than that it excelled in vocabulary and word meaning tests, in that the training that it received was apparently much more directly related to the latter.

Van Wagenen and Kelly (62) made an exhaustive study of the

inter-relations, as determined by the correlation method, of the abilities involved in theme writing, in reading for understanding of content, and in completing mutilated sentences (Trabue L and M), and the relation of these abilities to marks received in whatever academic courses had been taken in the sophomore year. The subjects were 98 sophomore students, 86 women and 12 men, in the University of Minnesota. All possible inter-correlations and numerous partial correlations are reported. It will be possible to mention only those bearing immediately upon the test under discussion. The two completion scales yield a reliability coefficient of 0.44 which is not high. The correlation coefficient for first semester marks and Trabue Scales L and M is 0.28, which is considerably lower than the coefficient of 0.41 found by Colvin for first semester freshmen grades. Other coefficients of correlation with the Trabue Scales are as follows:

Sophomore marks, both semesters	0.31
Rhetoric marks, both semesters	0.53
Scores in reading tests	0.25
Scores in Themes A and B	0.44

The homogeneity of the group of students may account in part for the small coefficients.

The completion test has not been used as extensively as the opposites test as a component of test series for measuring general intelligence. An examination of the same 29 group intelligence examinations previously discussed, shows that the completion test, regardless of its form, has been included in nine tests. The disarranged sentences test is not here considered one of the forms of the completion test although it is related to it more or less closely.

Summary. The completion method as devised by Ebbinghaus, and as developed by Trabue has been designated, as has the opposites test, a test of real intelligence, the best single measure of mental ability. It has likewise been said to differentiate between groups of varying intelligence. Some investigators have considered it as an exercise in judgment; as a form of reasoning test; as showing fundamental differences in thought processes; as a test of selective judgment. Depending somewhat upon the

method of administration and the nature of the elisions in the text, it has been held to measure verbal memory; to indicate a command of language; to show peculiarities of language development. By two investigators, it has been termed the best all-round measurement of language ability.

With elementary school children, age progression in performance has been shown. Fairly significant correlations have been obtained for small groups between intelligence and grades. It has also been shown that for elementary school pupils the completion test measures the same ability as that measured by such tests as opposites, vocabulary, written directions, practical judgments. The ability measured by this test is more closely related to that involved in English work than in mathematics, that is, the verbal factor is common to both.

With one group of college students the completion test has been found to correlate as well with grades as the entire Alpha test does, although it correlated less well than the vocabulary and analogies tests.

The different conditions under which the tests were given, the differences in time allowed, in the exact form of the test used, in the nature of the groups tested make a direct comparison between the results impossible. One investigator found the completion test more predictive of success in English than in mathematics, whereas another found it equally predictive of both. Here again, we are confronted with the necessity of studying the nature of the groups tested.

c. Vocabulary and information tests

The vocabulary test is listed by Whipple (65) as a test of "intellectual equipment." It is a test of what an individual knows rather than what he can do. He says,

> Since nearly all thought and expression is couched in linguistic form and since the intellectual progress of a child at school is, in a sense, a process of augmentation of his vocabulary, and of refinement in its use, it seems not unreasonable to assume that the determination of the size of his vocabulary will be of significance and value in estimating his general intellectual status.

Kirkpatrick (31) is equally convinced of the value of determining the size of an individual's vocabulary as an indication of his intellectual equipment. He says,

> The vocabulary of a person represents in a condensed and symbolic form all that he has experienced and imagined. The breadth of his mental experience is indicated by the number of words that have for him a meaning, while accuracy of his thinking is shown by the constancy and exactness of meaning with which he uses words. The studies of vocabularies ought, therefore, to be an important branch of psychological investigation.

Kirkpatrick was the first to suggest the possibility of approximating the size of an individual's "understanding" vocabulary by means of a comparatively few selected words. He selected a list of 100 words as follows: a word in a definite position, first, second, etc., on every sixth page of Webster's Academic Dictionary, containing 28,000 words on 645 pages was selected. The list of words thus obtained was given orally to groups of subjects, who were asked to check the words not known. Whipple (66) has proved that this method is somewhat inaccurate, in that it leads to the over-estimation of the size of the vocabulary. Kirkpatrick found as the result of testing about 2000 people ranging from the second grade through college, that there was a very definite age progression, although there were wide individual variations. Whipple (66) has confirmed these results.

Terman and Childs (53) prepared an entirely different list of one hundred words, based on 18,000 more commonly used words in Laird and Lee's Vest-Pocket Webster Dictionary, 1904 edition. This list of one hundred words is included in the Stanford Revision. The results of its application have shown a steady growth with age in vocabulary index.

Brandenburg (7) selected 200 words from Webster's Academic Dictionary with which he tested over 2000 pupils in 68 different classes, from the second to the twelfth grades, in six different school systems. The children were required to write sentences using each word correctly. He found an average gain from year to year, from grade two to eight, of approximately 1400 words.

Although a knowledge of the meanings of words and some facility in their use is presupposed by a large proportion of the tests in the Binet-Simon Scale (5), certain tests are included

in the scale specifically to test vocabulary. In the 1905 series, test 14, "Verbal definitions of known objects," requires definitions of concrete objects; and test 30, "Definitions of abstract terms," the ability to give differences between such words as weariness and sadness. In the 1908 scale, Binet and Simon recognize different grades of definitions given by children of different degrees of mental development. Thus, definitions by use only are given normally by six-year-olds; definitions superior to use by nine-year-olds; and definitions of abstract terms by eleven-year-olds. They found that giving of definitions, together with the arrangement of weights and the interpretation of pictures "are among the tests which are most frequently passed before age." They are less influenced by the child's surroundings than some other tests and are therefore more adequate expressions of spontaneous intelligence.

Again, they list six tests "uniquely expressive of intelligence" which may be considered as forming for the laboring class of Paris and the environs the borderline between moronity and the normal state. These tests are: 1, arrangement of weights; 2, answers to questions difficult of comprehension; 3, construction of a sentence containing three words; 4, definition of abstract words; 5, the interpretation of pictures; 6, the making of rhymes. The ability to handle abstract ideas is a particularly good indication of intellectual development. William James (24) has expressed this idea in the following picturesque way:

> Without abstract concepts to handle our perceptual particulars by, we are like men hopping on one foot. Using concepts, along with particulars, we become bipedal. We throw our concept forward, get a foothold on the consequence, hitch our line to this, and draw our percept up, traveling along with a hop, skip, and jump over the surface of life at a vastly rapider rate than if we merely waded through the thickness of particulars as accident rained them down upon our heads. Animals do this but men raise their heads higher and breathe fully in the upper conceptual air.

The three definitions tests are retained in the 1911 revision, with a shift of the abstract definitions to the twelve year group of tests.

In spite of the fact that language tests are considered valuable tests of intelligence, as such, Binet recognizes that environment

is a very potent force in the development of language. In reviewing the work of Decroly and Degand (17) Binet shows the superiority of the social status of the group of children studied by them, over the children in the primary schools of Paris. Of the thirteen tests in which Decroly's and Degand's children were a year and a half in advance of Binet's, language is a primary requisite in six. Thus, intellectual superiority manifests itself especially in tests in which language plays a part.

In the Stanford Revision of the Binet-Simon Scale, definitions tests, or vocabulary tests play even a larger part. Terman requires the giving of definitions in terms of use, and superior to use, each, of a younger group than Binet and Simon did; but retains the definitions of abstract terms in the twelve year group. Terman has also noted that the formation and use of abstract ideas represents, *par excellence*, the higher thought processes.

The vocabulary test, however, is considered the most valuable single test of the Stanford Revision. In an effort to meet some of the many criticisms that have been hurled at this test, Terman (52) made a special study of it with respect to its validity as a measure of general intelligence, the justification for its use with children of non-English speaking parents, and its relation to the ability to name words in a free association test. The correlation between the vocabulary test and the Stanford Revision mental age for a miscellaneous group of 631 subjects was 0.91. Terman considers this very significant. Since the members of the group are scattered from the first grade to the first year of high school, it is obvious that the elimination of the chronological age factor would greatly reduce the coefficient. There is another definite, although not exactly measurable, spurious element in the correlation, in that the vocabulary test as such, together with the other tests of definitions contributes to the mental age score. Only in so far as the Stanford Revision is conceded to be a valid measure of intelligence is the vocabulary test validated, in this particular study, and Terman seems to be somewhat over-enthusiastic when he says, "We believe it will be possible before long to measure the intelligence level almost as accurately by means of a vocabulary test of 100 crucial words as it can now be measured by any existing intelligence scale."

Terman (49) found a correlation coefficient of 0.487 between the number of words given in three minutes and size of vocabulary, for a miscellaneous group of 360 children; and a coefficient of 0.535 for word naming and mental age for a miscellaneous group of 480 children. Whipple (66) reports a coefficient of 0.53 for a group of 58 college students between Kirkpatrick's vocabulary index and the word building test. James (23) reports coefficients of correlation between Terman's vocabulary test and intelligence as measured by a test similar to the Army Alpha, and grades as follows:

	CASES	r
Vocabulary index and intelligence	82 juniors	0.55
Vocabulary index and intelligence	25 seniors	0.59
Vocabulary index and grades	76 juniors	0.51
Vocabulary index and grades	46 seniors	0.60
Intelligence and grades	91 juniors	0.52
Intelligence and grades	45 seniors	0.48

The vocabulary index correlates as well as the intelligence test with grades.

Colvin (16) found a correlation coefficient of 0.45 between a vocabulary test made up of twenty-five of the difficult words in Terman's list, and the first semester grades, for a group of 212 college freshmen, which is similar to the one found by James.

Carothers (12) reports an average coefficient of correlation of 0.00 between the vocabulary test (specified by Whipple's Manual) and other tests, including opposites, mixed relations, completion, word naming, word building, cancellation, etc., for a group of 100 college freshmen. The information test showed a similar lack of correlation. The only suggestion she offers for this curious finding is that "information and vocabulary differ from the other tests of the series in that they are indicative of one's learning rather than one's innate ability." It would seem, however, that opposites and mixed relations tests are just as indicative of one's learning. The vocabulary test correlates more highly with language grades than any other test, including opposites, mixed relations and completion. The coefficient of 0.41 is higher than any other coefficient of correlation between tests and grades that she found.

It would be difficult to explain the lack of correlation of this test with other verbal tests were it not that all of her coefficients are extremely low. The method of giving and scoring the tests may be held partly responsible.

Brandenburg (7) reports coefficients of correlation of his vocabulary index with language grades and average grades for groups of children in grades four through ten. The coefficients for language grades range from 0.54 to 0.90, with an average of 0.76. The coefficients for average grades range from 0.39 to 0.85 with an average of 0.63. The coefficients show a fair degree of consistency and while the groups studied are too small for generalizations, it is indicated that for school children there is a definite positive relationship between the vocabulary index and grades in school subjects.

All of the vocabulary studies thus far reported have been made in the "understanding" or reading vocabulary, *i.e.*, in the comprehension side of language ability. It would be interesting to devise a vocabulary test to measure the other phase of language ability, *i.e.*, construction. Instead of giving the meaning of a word, the subject would be given a definition to which the word that it defines must be supplied. This task would probably be a very much more difficult one.

The ability to comprehend language comes first in order of development. This fact has been noted by all investigators who have worked with very young children. Children give evidence of understanding what is said to them for an appreciable length of time before they are able to talk, or, with the majority of people, the ability to comprehend keeps ahead of the ability to construct and compose. To illustrate, it is generally accepted that an average individual's reading vocabulary is far in excess of his writing or speaking vocabulary. It is possible to comprehend style and precision in the writing and speaking of others without being able either to speak or write with equal facility. It would be of interest to determine whether the ability to use words is in a numerical ratio to the ability to understand words, for given individuals. It is strongly suspected that such is not the case.

The range of information test was devised by Whipple (63) as an

extension of the vocabulary test. The vocabulary test has likewise been termed a form of information test. The 100 test words in Whipple's list have been selected not by chance as the words are selected for a vocabulary test, but by careful consideration so that each will be representative of some special field of knowledge or activity. It is the selection of the test words that constitutes the real difference between the information and vocabulary tests. In the latter, unusual words and words infrequently met with outside a special subject predominate.

The writer is of the opinion that the information test is not essentially a language test in that facility in the handling of abstract verbal concepts is not the predominating factor. It is directly related to the vocabulary test, however, in that they both involve to an extent the ability to learn the names of things, events, etc., although, as it has been said, the information test deals more with a specialized and less abstract type of knowledge than does the vocabulary test.

A brief discussion of the relations of this test to general intelligence and to other tests will be included.

Results obtained by Whipple and Smith (44) are summarized in Whipple's Manual (65). There is an apparent increase in the ability to define or explain technical terms with increased maturity. No other published records of the use of this test appear until that of Bell (2) in 1917. Bell combined the data gathered by Miss Smith with additional data of his own, making a group totaling 596 students. The words in the test were then arranged in a decreasing order of difficulty, on the basis of these 596 scores. Decile distributions for the four college classes are given, which show an increase in score with increase in years in college at almost every decile. Bell has also grouped the words according to nine various fields of knowledge represented. His classification seems quite arbitrary in many instances, although the difficulty in making such a classification is recognized. Items pertaining to history and literature are more generally familiar than those pertaining to the biological sciences.

King and M'Crory (30) report coefficients of correlation of the Whipple information test with other tests and university grades as in the following table on page 43:

	276 WOMEN	268 MEN
Arithmetic speed (Courtis)	0.06	Neg.
Arithmetic accuracy (Courtis)	0.13	Neg.
Completion (Simpson)	0.21	Neg.
Opposites (Simpson)	0.24	0.56
Analogies (Whipple)	0.28	Neg.
Visual imagery (painted cube)	0.23	Neg.
Logical memory (Whipple)	0.31	0.13
Test average	0.60	0.65
University grades	0.41	0.44

Whereas there is but slight correlation with other tests, there is a fair degree of correlation with university grades. Although it is less than that between opposites and grades, it is greater than that between analogies and grades.

Uhl (60) reports the following coefficients of correlation for a group of 100 university students between the Whipple information test and other tests and grades as follows:

Trabue Completion Scale K	0.18
Trabue Completion Scale M	0.42
Opposites	0.26
English grades	0.38
Mathematics grades	0.43

These coefficients are small as are all of those reported by Uhl.

It has already been reported that King and Gold (29) with a group of 50 college students found a higher degree of correlation between Whipple's information test and both easy and hard opposites than between vocabulary and opposites. They report a coefficient of 0.45 for the correlation of information and vocabulary.

The average zero coefficient found by Carothers (12) for the correlation of information and all other tests has already been discussed.

Two of the 31 sub-tests studied by Jordan (25) (26) were information tests, Alpha 8 and Terman 1. He found that Terman information correlated best of any sub-test with average grades, and third highest with the composite score based on all tests.

With no other criterion, however, does the information test correlate sufficiently well to be included among the five highest correlations.

Although statistical data are not abundant to prove the validity of the information test as a test of intelligence, it has been included in a large proportion of the intelligence examinations. Of the 29 intelligence tests that have been mentioned, 14 include tests of general information, and 5 vocabulary tests, designated as such. Thurstone (55) included it in both editions of his "Psychological examination for college freshmen and high school seniors." He says,

> Although the general information test is not a direct test of intelligence, it is an excellent indirect test of that attribute. Other things being equal, it is safe to assume that the bright person will acquire unwittingly a greater range of information than the mentally less gifted person.

Thorndike (46) who has likewise included an information test in his series, says that a straightforward information test is a valuable element in tests designed to measure intelligence. The criticism that information tests measure specific training rather than innate ability is met in part by Whipple (64), who in a discussion of the relative importance of training versus endowment as factors conditioning success in intelligence tests, concludes that

> Heredity to some extent makes its own environment. . . . . Some considerable part, therefore, of the better scores made by . . . . children which are *immediately* attributable to their superior training must, after all, be *ultimately* ascribed to their better endowment.

Summary. Both the size of vocabulary and the nature of definitions given for concrete and abstract terms have been considered as diagnostic of general intellectual status, and, as a result, tests to measure these two related abilities have been included in general intelligence test series. A number of investigators have shown the steady increase in size of vocabulary with age. Because of the relatively high correlation with the Stanford-Binet mental age, one investigator has been led to suggest the possibility of devising a vocabulary test which will be as predictive of the level of intelligence as any existent intelligence scale.

The coefficients of correlation of vocabulary indices with grades of elementary school pupils are relatively high. With language

grades alone, they are still higher. On the whole, fairly good correlation has been shown to exist for groups of college students, between vocabulary indices and grades, total intelligence score, and for one group, a word building test.

The information test, which has been classified as a form of vocabulary test, or at least as closely related to the vocabulary test, has been considered a valuable element in general intelligence tests, although data are not abundant to show that it correlates very highly with criteria such as grades and estimated intelligence. It would seem to be less useful as a single test for measuring intelligence than opposites, completion, or analogies.

It is suggested that a vocabulary test, designed to measure the constructive phase of language ability, would be of value in as much as there seems to be a difference between ability to understand and to use words, and since the existent vocabulary tests measure chiefly the understanding or comprehension phase of language ability.

d. Analogies test

The analogies test, listed by Whipple (65) as a test of controlled association has been variously considered as follows: 1, by Woodworth and Wells (70), who first attempted to use the analogous relation as a mental test, as a measure of "flexibility of mental performance" and "skill in handling associations;" 2, by Burt (11), who first used the term "analogies" in connection with the test, as involving "perceptions, implicit or explicit, of the relation and reconstruction of the analogous one by so-called "relative suggestion;" 3, by Briggs (8), as an indication of "rapidity and accuracy of mental reorientation;" 4, by Bickersteth (4), as a test of reasoning power; 5, by Whipple et al. (67), as demanding "the perception of relatively abstract verbal relationships," and as "one of the best indices of this important aspect of general mental ability," as bringing out "an ability that is decidedly symptomatic for the purposes of selecting gifted children."

These experimenters have generally found a high degree of reliability and a fairly high positive correlation between the analogies test and estimates of intelligence and with other single tests of general mental ability. Burt (11) in 1911 found a reliability coefficient of 0.90 and a coefficient of correlation with estimated

intelligence of 0.52 for a group of 60 to 75 boys and girls, aged 11.5 to 13.5 years. Wyatt (72) in 1913 found a correlation between estimated intelligence and the analogies test of 0.80 for one group of 34 boys and girls, aged eleven and thirteen, and of 0.62 for a second group of 41 girls, aged ten to twelve. He considered these coefficients as "remarkably high." Bickersteth (4) in 1917 found a consistently high reliability for this test with four age groups, 10.5 through 13.5 and for groups of boys and girls separately.

Since Jordan's two studies (25) (26) have been reported in some detail in the discussion of the synonym-antonym test, they will be briefly summarized with respect to the analogies test, at this point. All four of the group tests which he used included analogies tests, Alpha 7, Otis 7, Terman 7 and Miller 3. Among the correlations of the 31 sub-tests with various criteria for a small group of high school pupils, the analogies test stood as follows: third highest with combined grades, with Terman information and Alpha mixed sentences standing first and second, respectively, and the opposites tests sixteenth, seventeenth, and twenty-second; third highest with English grades, which is not as high as the synonym-antonym test; fourth highest with teachers' estimates of intelligence, which is less high than the synonym-antonym test. It correlated considerably less well with the Stanford Revision mental age and with a composite of the four tests than did opposites. It does not appear among the four sub-tests which show the highest average correlation with all of the criteria.

King and M'Crory (30) found the following coefficients for groups of university students:

	276 WOMEN	268 MEN
Arithmetic speed (Courtis)	0.17	0.04
Arithmetic accuracy (Courtis)	0.22	0.18
Completion (Simpson)	0.46	0.58
Opposites (Simpson)	0.52	0.77
Information (Whipple)	0.28	Neg.
Visual imagery (painted cubes)	0.20	Neg.
Logical memory (Whipple)	0.32	Neg.
Test average	0.72	0.74
University grades	0.14	0.40

The opposites test correlates more highly with analogies than any other single test. Carothers (12) also found a comparatively high degree of correlation, 0.57 between opposites and mixed relations or analogies, and a good correlation, though somewhat lower, between completion and mixed relations, 0.48.

Baum, Litchfield and Washburn (1) found a correlation coef ficient of 0.39 between analogies and academic rank for a group of 31 college seniors, previously described. The coefficient of correlation between a combined score for an opposites and an analogies test was 0.40, which is significantly larger than the coefficient for the analogies test taken alone. With their second group of 50 students, 25 from the upper and 25 from the lower end of the senior class, they found that the analogies and opposites tests served to separate the upper from the lower group.

Colvin (16) found for a large group of university students that an analogies test of ten items correlated more highly with first semester grades than either a vocabulary or sentence completion test; as high (0.52) as an average of the two Brown University series of tests with grades (0.53); higher than the Army Alpha with grades (0.43); and as high as a combined score for the Army Alpha and the Brown University test with grades (0.51).

For university students, the coefficients of correlation of an analogies test with grades are lower than those found for grammar school pupils.

Pintner and Renshaw (40) report a coefficient of 0.785 for 52 second-year normal college students between an analogies test and the Otis Intelligence Examination, Form A.

The analogies test has been accorded a prominent place in many intelligence examinations. In the Army Alpha, it was found to give results on a par with the synonym-antonym test. The distributions of scores are very similar for the two tests, in that they both show a large proportion of zero scores. It was the most effective test in the whole series for differentiating officers from enlisted men (9). It has also been incorporated in the Otis Group Intelligence test, the Thurstone Tests for college freshmen and high school seniors, in the 1920 and IV editions; in the Thorndike Examinations, and in other series, appearing in twelve out of twenty-nine series of intelligence tests.

Two attempts at standardization of the analogies test appeared in 1920, one by Pintner and Renshaw (40), and a second one by Van Wagenen (61). The former is a standardization of 200 analogies, based on the results obtained from 917 subjects. The 200 analogies are arranged according to difficulty, with the percentile, sigma and point value, and the correct response as in the following illustration:

Point value: 10.0
Percentile: 62
Sigma = 3.76 heat : a gas :: cold : ice winter water refrigerator

Van Wagenen has devised four equivalent lists of 50 analogies, based on the results of experimentation with 142 sixth and seventh grade pupils and 122 college juniors and seniors. The intercorrelations between the four tests range from 0.83 to 0.89. The analogies are presented in the following form:

color : red :: name : ————

In general then, the analogies test has not been as unreservedly acclaimed a test of general intellectual ability as have the opposites, completion and vocabulary tests. It has been considered, however, as a valuable index of reasoning power and of the ability to perceive relatively abstract verbal relationships. As such, it has been useful in separating gifted from normal or subnormal children. In as much as success in the analogies test is conditioned by an adequate understanding of the stimulus words, it may be considered a language test.

With the groups of elementary school children it has been shown to have high reliability and to correlate well with estimates of intelligence. With a high school group, the analogies test was more predictive of average grades than the synonym-antonym test but less so than the Terman information and the Alpha mixed sentences tests. It correlated less well with grades in English, with teachers' estimates of intelligence, with a composite score and showed a lower average correlation with all criteria than did the synonym-antonym test. In the army testing, it was as successful as the synonym-antonym test in differentiating officers from enlisted men.

For university students, the coefficients of correlation with average grades are lower than for elementary school groups. It has been found to correlate with grades as well as the average of the two Brown University series of tests, better than the entire Army Alpha, and as well as a combined score of the Army Alpha and the Brown University tests.

IV. GENERAL SUMMARY OF THE STATUS OF TESTS WHICH MEASURE LANGUAGE ABILITY DIRECTLY OR INDIRECTLY

Although an attempt has been made to summarize the development and uses of some specific tests and although certain tendencies have been indicated on the basis of the conclusions offered by a number of investigators, it is with a certain reluctance that one attempts to make a further summary of all of the tests in their relation to each other. The farther away one gets from the experimental data, the easier it is to generalize. It is easy at this point to say, for instance, that opposites, vocabulary and completion tests are excellent tests of general intellectual ability and they may be, but one is forced to recognize the weakness of much of the data on which such a statement is based. Correlation has been the principal method of validation and in many instances slight notice if any has been taken of the spurious correlation due to the heterogeneity of the group, or to identical or common elements in the criteria. Thus, the age factor, which is a most vital one in all studies of children has frequently been overlooked, so that one suspects that the coefficients derived are "too high;" again it is asserted that a given test is practically as efficient a prognosticator of intellectual ability as a series of tests, because of its correlation with the series, despite the fact that the test series is heavily weighted with the score derived from the one test. Correlations between tests and grades, and estimated intelligence are consistently higher for elementary school pupils than for college students, which may be due to the homogeneity of the college group rather than to a greater real correlation in the case of the elementary school pupils.

A more fundamental difficulty in evaluating the results of experimentation with specific tests and in relating them to each other lies in the varied nature of the test material, in differences

in conditions under which tests are given, including actual working time allowed, and in scoring. In order to have any certainty as to the actual relations of tests to each other, their common elements and their relations to given criteria, one would have to apply the same tests under the same conditions to large groups of individuals of different ages, who had had equal environmental opportunities; uniformity in the treatment of results would have to be insisted upon; and conclusions based on correlational data made with full realization of the possible causes of spurious correlation.

With recognition, then, of the nature of much of the evidence, we can perhaps indicate some general tendencies. The opposites, vocabulary and completion tests have all been considered good measures of general mental ability, and the analogies test somewhat less so. In at least some forms of the tests, progressive age differences have been demonstrated and they have been found to differentiate between groups of varying intellectual development. Correlations indicate that they are fairly predictive of average grades, more so of English grades, and less so of mathematics grades.

There is evidence that these tests have a factor in common, which is, obviously their verbal nature. But it remains to be determined to what extent the ability to use and understand words is the only factor in common. Kelley (27) has asserted that for elementary school pupils at least the tests listed above, do not measure disparate functions at all. It has not been proved that they do or do not measure the same function in adults. It would be possible by the method of partial correlation to determine, at least in part, to what extent a facility in understanding words or the size of vocabulary conditions the scores in an opposites or analogies test. A better procedure might be to relate the performance in an opposites or analogies test in which the stimulus words are known to be difficult, to a subsequent performance in defining the terms that have been used in the first test.

If it is conceded that the relation of language to perceptual and thought processes is such that the development of one is conditioned by the development of the other, then it may be that the stage of development of one, *i.e.*, language, will indicate the

development of the other. It may be, then, that the most dependable tests of "intelligence" will be those that involve language as a primary factor. In their contributions to the symposium on "Intelligence and Its Measurement" conducted in the *Journal of Educational Psychology* (46), Terman and Haggerty seem to think it possible and probable that the best tests of intelligence will be those which involve the use of language or other symbols, such as the completing of analogies, naming opposites, matching proverbs, understanding difficult passages, completing elliptical sentences, and also solving arithmetical problems. On the other hand, Thorndike, Pintner, Colvin and Henmon form a group which contends that whereas the language tests are fairly predictive of a limited sort of intelligence, intelligence in a broader sense should imply other abilities as well. In other words it would seem that intelligence has been defined in terms of certain abilities closely allied to facility in the understanding and use of words, and it follows naturally that tests involving a similar ability will correlate with it. This statement is not supposed to imply that the opposites, vocabulary, completion, and analogies tests do not test abilities distinct from that of handling verbal concepts, but it remains to be demonstrated to what extent they are tests of "language ability."

CHAPTER II

General Intelligence Examinations at the Johns Hopkins University

The data analyzed in this section are drawn from the results of Intelligence tests administered to the entering college and engineering students in The Johns Hopkins University. The Thurstone, 1920 edition, was given to entrants in October, 1920; the Thurstone IV was given to the next entering class in October, 1921. The Anderson test was given in February, 1922, to the class in English composition which included most of the men entering the previous October who were still in the University. In October, 1922, the class then entering was given the Anderson test and some additional specific tests. The first form of The Johns Hopkins Combination test was given to the entrants in October, 1923.

1. Thurstone Psychological Examinations, Test IV and 1920 Edition

The Thurstone "Psychological examinations for college freshmen and high school seniors" both Test IV and the 1920 edition are made up of six types of test material arranged in cycle-omnibus form. Although Thurstone (55) has pointed out that this type of test is not particularly useful when the diagnostic value of each part is being investigated, he considers it for administrative purposes "far superior to the separate giving of the six tests." In view of the relatively low degree of correlation with college grades, for example, 0.29 for first semester marks of a group of 650 students, 0.33 for first year college marks of a group of 278 Vassar students (48), the question arises whether the form of the test or the actual types of test material included in the whole test is at fault.

Despite Thurstone's warning, it was thought that a study of the individual types of tests included in his examination might throw some light on the question of the relationship of the different

types of tests and indicate individual differences in performance for types of test material, whether the more verbal, informational, or mathematical type.

"Test IV," which was issued a year before the 1920 edition, is made up of the following types of test material: 1, information; 2, analogies; 3, sentence completion; 4, syllogisms; 5, proverbs; 6, number completion. There are twice as many test items of information and analogies as of the other types.

The information test items are in the multiple response form, that is, the subject selects one response as correct out of four possible responses. An examination of the items in the information test shows that it is not essentially a form of the vocabulary test, in that only 16 of the 40 information items could be considered as possible stimulus words for a vocabulary test, and among these 16 there is a preponderance of technical words, such as tedder, slice, piccolo, kilowatt, etc. A comparison of the items with those in the Iowa High School Content Examination (22), which as its name suggests, is designed to cover information acquired directly from high school curricula, indicates that the Thurstone information is largely drawn from extra-curricula activities, from current events, sports.

The analogies test is also of the multiple response type, the subject being required to underline two of five possible words which bear a certain relationship to each other. All of the stimulus words used are probably in the vocabularies of high school graduates and college freshmen, since they are all relatively simple. Therefore, an understanding of the words dealt with in the test, although essential to success, is probably not the primary factor conditioning success.

All of the items in the completion test are taken from the Trabue "Language Completion Scales." The elisions are such that no technical or special knowledge is required of the subject but rather, an ability both to understand and to construct somewhat simple language forms. No partial credits are allowed in scoring, although the number and difficulty of the elisions vary in different sentences.

As is true of the analogies and completion tests, the syllogisms test does not seem to require a knowledge of words beyond that

of a high school graduate. It does require, however, an ability to handle abstract verbal concepts.

The proverb checking test is in fact a reading test and will be designated as such. In at least some of the items of the test, the subject is required to abstract the meaning from a difficult, though short prose passage.

There is only one test of the six types that is entirely non-verbal, *i.e.*, the number completion test. The subject is required to complete number series without the suggestion offered by the multiple response form of test.

We have analyzed the responses to each item in the Thurstone IV test given by a group of 197 college freshmen and derived a score for each subject in each of the six tests in the series equal to the number of correct items, omitting the examples which are included in the test. Since most of the subjects did not finish the test, it is obvious that, owing to its arrangement in cycle form, one more item (or two more in the case of the information and analogies items) of one specific type may be responded to than in the remaining types.

The range, average, and standard deviation of the scores in the six tests are given in table 1. The analogies test contributes more, absolutely and relatively than any other test, and the syllogisms least. On the whole, however, it would seem that the test material of all six types is of approximately the same difficulty.

Coefficients of correlation[1] for each of the six tests with the total score and with each other test are shown in table 2. The range of coefficients of correlation with the total score is from 0.915 to 0.824 which indicates a decided tendency for one test to contribute about as much to the total score as any other test.

[1] All correlations have been derived according to the method described by Toops (57) which involves the Pearson formula for gross measures, adapted for use with plotted measures. The formula is as follows:

$$r = \frac{\frac{N}{2}\left[(\Sigma x^2 + \Sigma y^2) - \Sigma(x - y)^2\right] - \Sigma x \cdot \Sigma y}{\sqrt{N \cdot \Sigma x^2 - (\Sigma x)^2}\ \sqrt{N \cdot \Sigma y^2 - (\Sigma y)^2}}$$

Partial correlation coefficients have been derived by Yule's formula (75):

$$r_{12.3} = \frac{r_{12} - (r_{13} \cdot r_{23})}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}}$$

The inter-correlations of the six tests range from 0.655 for number completion with reading to 0.770 for syllogisms and analogies. Here again the range is fairly short. It seems clear, then that, in general, the subjects do about as well with one type of test

TABLE 1

Range, average, and standard deviation of scores in the six tests of the Thurstone IV series, and of the total score for 197 entering students

NAME OF TEST	RANGE	AVERAGE	S.D.	POSSIBLE SCORE
Thurstone IV, total score	32–138	86.27	22.19	160
Information	5– 37	21.72	6.06	40
Analogies	2– 39	24.16	6.58	40
Sentence completion	2– 17	10.00	2.92	20
Syllogisms	0– 17	9.98	3.12	20
Reading	0– 20	10.80	3.41	20
Number completion	0– 19	11.81	2.99	20

TABLE 2

Coefficients of correlation between total score and tests in the Thurstone IV series, for a group of 197 entering students

	TOTAL SCORE	INFORMATION	ANALOGIES	SENTENCE COMPLETION	SYLLOGISMS	READING	NUMBER COMPLETION
Total score		0.888*	0.915	0.824	0.859	0.852	0.837
Information	0.888		0.749	0.678	0.678	0.684	0.728
Analogies	0.915	0.749		0.720	0.770	0.746	0.730
Sentence completion	0.824	0.678	0.720		0.710	0.714	0.690
Syllogisms	0.859	0.678	0.770	0.710		0.761	0.707
Reading	0.852	0.684	0.746	0.714	0.761		0.655
Number completion	0.837	0.728	0.730	0.690	0.707	0.655	
Average r (omitting total score)		0.703	0.743	0.702	0.723	0.712	0.702

*Probable errors are ± 0.02 or less.

material as another in this form of test. Or, in other words, as Thurstone has already stated, the cycle form of test does not lend itself to the sort of analysis that is being attempted. If the coefficients of correlation of each test with each other test are averaged, the analogies test shows a slightly higher average.

With full recognition of the fact that the similarity between the inter-test correlation coefficients, due apparently to the form of the test, hardly justifies it, an attempt has been made to discover any special relationships between tests, by the partial

TABLE 3

Coefficients of partial correlation of the sentence completion test, with each of the remaining tests in the Thurstone IV Examination with each test constant. 197 cases

C, sentence completion; *A*, analogies; *I*, information; *S*, syllogisms; *R*, reading; *N*, number completion.

TESTS	*r* OF ZERO ORDER	TEST HELD CONSTANT	*r* OF FIRST ORDER
C. and I	0.678*	A.	0.302*
		S.	0.379
		R.	0.371
		N.	0.354
C and A	0.720	I.	0.436
		S.	0.385
		R.	0.402
		N.	0.491
C. and S	0.710	I.	0.455
		A.	0.351
		R.	0.367
		N.	0.434
C. and R	0.714	I.	0.423
		A.	0.359
		S.	0.372
		N.	0.425
C. and N	0.690	I.	0.390
		A.	0.347
		S.	0.377
		R.	0.420

*Probable errors are ± 0.05 or less.

correlation method. In table 3 are shown coefficients of partial correlation for the sentence completion test with each of the other five tests, each of the remaining four tests being held constant in turn. It is impossible to make deductions of any value from

these coefficients of the first order. Nothing is shown that is not already indicated by the average of the coefficients of each test with each other, namely, the slight tendency for the analogies test to have most in common with the other tests; syllogisms and reading, less; and information, sentence completion, and number completion least of all.

It is very difficult, if not impossible, to translate the results reported thus far into terms of language factors. One might be tempted to make a speculative statement to the effect that the analogies test shows the highest average because of the language factor which it has in common with the sentence completion and other tests, on the one hand; and its relation to the other so-called reasoning tests, such as syllogisms and reading, on the other hand, were it not for the fact that in the 1920 Thurstone Examination, the stituation is exactly reversed. That is, the analogies test shows the lowest average correlation, despite the language and reasoning factors which it still has in common with the other tests.

One further attempt has been made to determine the relations of the tests by means of correlation with grades in college courses.

Anyone attempting to use grades as criteria by which to measure the predictive ability of tests is confronted by the problem of the unreliability of grades, due partly to the subjectivity and variability of standards employed in different college courses resulting in inaccuracies in grading; partly to variation in attitudes of students towards grades, such as the indifference of able students, or the persistence of less able students; and finally, to the effects of extra-curricular activities, economic pressure, etc. It is impossible to control this second and third group of factors as well as to estimate the extent to which they affect grades assigned. Attempts have been made, however, to make some sort of estimate of the variability of grades due to differences in standards of marking.

Evidence of the inaccuracy of marks has been summarized by Carothers (12) and more recently by Wood (69) and need not be reviewed here.

It is obvious that if students are allowed fairly wide latitude in selecting courses of study, and if it is found that the average

grade assigned in some courses is higher or lower than that assigned in other courses, the averages made by students will be affected by this factor. That is, a student who happens to take a preponderance of courses in which the average grade assigned is relatively low, will automatically make a lower average grade than the student in courses in which higher grades are assigned, although the two students may be doing equally good, or bad work.

In order to show, in part, the variability of the grades upon which this study is based, we have made distributions of all of the first semester grades in every college subject in the curricula of three groups of students included in this study, that is, for the freshmen entering The Johns Hopkins University in 1922, and 1923, and those men in the group entering in 1921 who remained in the University until February and were members of an English composition class.

An examination of the range, average, and standard deviation of the grades in each course has shown at once the wide variation in types of distributions and in average grades assigned.

There is a tendency for the grades to average higher in small classes. The variation in the size of the average may be due to one or all of several factors. Standards of grading may differ so that one instructor will assign lower or higher average grades than another for the same work in subjects of equal difficulty. That subjects are of varying degrees of difficulty is apt to be true. It is also true that the selection of students in a given class varies from year to year, so that if for example the group is poor, the average grade should justly be lower than that assigned for another group composed of a better grade of student.

None of the above-mentioned factors influence a coefficient of reliability of average grades (determined by correlating the average grades of two semesters) inasmuch as they do not condition an individual grade at different times of the year. Therefore our reliability coefficient of 0.765 for 150 cases, while fairly high, tells us nothing of the effects of those particular factors.

The number of credit hours varies in different courses. An ideal average grade should equate in some way the differences in credit hours, distributions, and the averages of grades in particular subjects. Such an average grade is purely hypothetical at present.

The method of deriving average grades employed in this study is as follows: The grade in each subject is weighted by the number of credit hours of the subject and these weighted grades averaged. To illustrate, if a subject makes 7.5 in English Composition (6 credits), 8 in French Elements (8 credits) 6 in Physics (12 credits) and 9 in Public Speaking (2 credits) his average grade will be determined as follows:

$$\begin{array}{rcl} 6 \times 7.5 & = & 45 \\ 8 \times 8 & = & 64 \\ 12 \times 6 & = & 72 \\ 2 \times 9 & = & 18 \\ \hline 28 &) & 199 \end{array}$$

7.10, average grade

Although this method of averaging the grades eliminates the factor of credits making it impossible, for example, for a high mark in a two point course to offset a low mark in a twelve point course, we have not eliminated the possibility that a four point course in one subject may require more of a student than an eight point course in another subject, granted that he is able to do the work in both subjects equally well.

The correlation coefficients for Thurstone scores with college grades are given in table 4. All of the tests with the exception of number completion show very nearly the same degree of correlation with the average of two semester grades in English composition, although the correlation with the sentence completion is by a small degree the highest. Thus, as could have been predicted, the tests involving language correlate more highly with the ability required in an English Composition course, than does a strictly non-verbal test. The coefficients of correlation with the average of all grades in beginning courses in the laboratory sciences are all lower than those with English, due no doubt in part to the averaging together of grades in different subjects, such as physics, biology, etc.

The analysis of grades assigned in different subjects shows that the range, distribution and average of grades differ markedly in different courses. It follows that averaging grades of different combinations of subjects in which the grading is widely varied tends to make grades an unreliable measure. The small degree

of correlation may be due also to the fact that tests of the type included may not predict grades in science as accurately as they predict grades in English. There is less difference between the correlation of number completion with science grades and the correlations of the other tests with science grades, than in the case of the correlations with English grades, although it is again relatively very low.

The most obvious conclusion is that a cycle-omnibus test does not lend itself to a satisfactory analysis of its constituent parts. Although an examination of the test material, especially in the light of the historical account of three of the types of tests, would

TABLE 4

Coefficients of correlation between the Thurstone IV Examination, total score and score in each of the six component tests with average grades of at least two semesters in English Composition and in Science I

	ENGLISH COMPOSITION—AVERAGE GRADES OF 2 SEMESTERS (152 CASES)	SCIENCE I—AVERAGE GRADES OF AT LEAST 2 SEMESTERS IN BIOLOGY, CHEMISTRY, PHYSICS, GEOLOGY (155 CASES)
Thurstone IV	0.442 ± 0.04	0.223 ± 0.05
Information	0.401 ± 0.04	0.116 ± 0.05
Analogies	0.403 ± 0.04	0.188 ± 0.05
Completion	0.479 ± 0.04	0.193 ± 0.05
Syllogisms	0.418 ± 0.04	0.277 ± 0.04
Reading	0.438 ± 0.04	0.331 ± 0.04
Number completion	0.299 ± 0.05	0.179 ± 0.05

lead one to conclude that the test as a whole is heavily weighted with a language factor of some sort, no suggestion is offered, based on the result of this analysis as to the precise nature of the language factor.

The 1920 edition of the Thurstone Examination differs from Test IV in form in the reduction of the number of information and analogies items so that there is an equal number of each type of test items; in the substitution of the extra word test for the sentence completion test; and in the substitution of arithmetic problems for the number completion test.

The extra word test is more nearly like the vocabulary test

as such than any other test in the two Thurstone series, although it involves in addition to an understanding of the meanings of discrete words, the ability to note an element of similarity between a group of objects not present to sense, or between a group of abstract words. If the stimulus words are "dog bird cat house mouse," it is obvious that the subject will have no difficulty with the meaning of the stimulus words; but if they are "clarify explain argue illuminate elucidate," or "dissection dissolution integration destruction annihilation," the test is just as obviously a vocabulary test.

There are few items in the information test of the 1920 edition as in Test IV which are drawn from the knowledge required in subjects in the curricula of high schools. They show, rather, a preponderance of names of men prominent in current affairs, and of trade names, that is, of knowledge acquired incidentally. The test would seem then to be less closely related to the vocabulary test.

We have made an analysis similar to that of Test IV of the responses to each of the items in the test for 150 college freshmen. The range, average, and standard deviation of each of the six tests are given in table 5 and coefficients of correlation between each test with the total score and with each other test in table 6.

The correlations of each test with the total score are very similar and relatively high, ranging from 0.804 to 0.864. The intercorrelations range from 0.583 for analogies with information to 0.799 for extra word with information. (There is a slightly greater difference between the tests than in Test IV.) If the coefficients of correlation for each test with the other tests are averaged, the extra word test shows the highest average coefficient although it is but slightly higher than the average coefficients for information and arithmetic with the other tests. Analogies has the lowest average coefficient of correlation, whereas it had the highest average in Test IV. It is not possible from an examination of the stimulus words in the two tests to suggest the cause of this difference.

Although an examination of the test material might suggest that the 1920 edition of Thurstone's test is also heavily weighted with a language factor, particularly the understanding of language,

still arithmetic reasoning yields a higher average coefficient of correlation than syllogisms, reading, and analogies. The fact that the extra word test, which is a sort of vocabulary test, shows the highest average coefficient does not justify an inference that

TABLE 5

Range, average, and standard deviation of total score and scores in the six tests in the Thurstone Psychological Examination, 1920 edition, for a group of 150 entering students

NAME OF TEST	RANGE	AVERAGE	S.D.	POSSIBLE SCORE
Thurstone, 1920 edition, total score	28–156	83.5	23.2	174
Information	5– 28	14.5	4.3	29
Analogies	0– 28	12.6	5.5	29
Extra word	3– 25	16.2	4.0	29
Syllogisms	0– 27	13.7	4.1	29
Reading	0– 28	12.8	4.9	29
Arithmetic	2– 23	13.6	4.1	29

TABLE 6

Coefficients of correlation between the sub-tests and total score in the Thurstone Psychological Examination, 1920 edition, for a group of 150 entering students

	TOTAL SCORE	INFORMATION	ANALOGIES	EXTRA WORD	SYLLOGISMS	READING	ARITHMETIC
Total score		0.864*	0.804	0.856	0.837	0.828	0.860
Information	0.864		0.583	0.799	0.663	0.676	0.725
Analogies	0.804	0.583		0.591	0.586	0.627	0.626
Extra word	0.856	0.799	0.591		0.716	0.613	0.734
Syllogisms	0.837	0.663	0.586	0.716		0.681	0.698
Reading	0.828	0.676	0.627	0.613	0.681		0.618
Arithmetic	0.860	0.725	0.626	0.734	0.698	0.618	
Average r (omitting total score)		0.689	0.603	0.691	0.669	0.643	0.680

* Probable errors are ± 0.03 or less.

it is due to the language factor in common with the other tests inasmuch as its highest single coefficient of correlation is with information, the second highest with arithmetic, and the lowest with reading and analogies.

Coefficients of correlation with grades in college subjects are shown in table 7. They show low correlations with average grades for two semesters of English composition, with the highest coefficient between grades and analogies and the lowest between grades and arithmetic. The order is reversed for a fairly small group in English literature. That is, the highest coefficient is between grades and arithmetic and the lowest between grades and analogies. This order is retained for correlation with grades in beginning courses in science.

TABLE 7

Coefficients of correlation between the Thurstone Examination, 1920 edition, total score and score in each of the six component tests with average grades of two or more semesters in English Composition, English Literature, and Science I

	ENGLISH COMPOSITION AVERAGE OF TWO SEMESTERS (130 CASES)	ENGLISH LITERATURE—AVERAGE OF 2 OR MORE SEMESTERS (71 CASES)	SCIENCE I, INCLUDING BIOLOGY, CHEMISTRY, PHYSICS, GEOLOGY—AVERAGE OF 2 OR MORE SEMESTERS (137 CASES)
Total number right	0.400 ± 0.04	0.395 ± 0.06	0.337 ± 0.05
Information	0.395 ± 0.04	0.463 ± 0.06	0.326 ± 0.05
Analogies	0.422 ± 0.04	0.237 ± 0.07	0.218 ± 0.05
Extra word	0.304 ± 0.05	0.353 ± 0.07	0.257 ± 0.05
Syllogisms	0.323 ± 0.05	0.328 ± 0.07	0.318 ± 0.05
Reading	0.308 ± 0.05	0.278 ± 0.07	0.309 ± 0.05
Arithmetic	0.245 ± 0.05	0.501 ± 0.05	0.360 ± 0.05

The size of the coefficients together with the size of the group and the unreliability of grades, due to factors already described, combine to make generalizations from these data unwarranted.

To recapitulate, the cycle-omnibus form of test does not lend itself to an analysis of its constituent parts which might lead to conclusions as to the relationship of the parts. It has not been possible by inter-test correlations to analyze out factors common to certain tests but not to others which might be considered language factors. The fact that certain tests such as sentence completion and analogies tend to correlate more highly with English composition grades than with science grades, whereas

arithmetic reasoning correlates more highly with science grades than with English composition grades, indicates a differentiation between these tests as to the abilities that they measure. This is not a new finding but is mentioned as confirmation of the results of other investigators.

It is interesting to note that the highest single coefficient is that of the correlation of arithmetical reasoning with English literature, which indicates that the abilities involved in English composition and English literature are not identical. Wood (69) has recognized this fact and has said, "Between courses in English there seems to be no common denominator beyond some kind of concern with books." He found a coefficient of only 0.36 between two semesters of work in English (including courses in English literature) for a group of 170 students in Columbia University.

Correlations between test scores and grades for groups of engineering and arts and sciences students, separately and together

The relation of the Thurstone tests as wholes to average grades has been studied for engineering and arts and sciences students, separately and together (table 8). The average grades[2] of engineers are weighted more heavily with mathematics and sciences courses, while those of the arts and sciences group are made up more largely of non-mathematical and non-science courses.[3]

The coefficient of correlation between average grades and total test score is slightly higher in the case of the group tested by the Thurstone IV Examination, and slightly lower for the group tested by the Thurstone 1920 edition examination, for the engineering than for the arts and sciences students. The differences are not so marked as in the case of the Anderson Examination. In each case the average test scores are slightly lower for the engineering than for the arts and sciences students. In spite of the seemingly large proportion of language tests, the Thurstone Examinations are not specifically diagnostic of the ability involved in non-mathematical subjects, as opposed to that involved in mathematics and science.

[2] The method of deriving average grades is described in the section pertaining to grades.

[3] A more detailed description of the courses pursued by the two groups is given in connection with the discussion of the Anderson Examination.

II. ANDERSON PSYCHOLOGICAL EXAMINATION

The Anderson Psychological Examination, Forms I and II, consists of four sub-tests: 1, arithmetic; 2, synonym-antonym; 3, completion; and 4, information—a description of which follows.

The arithmetic test consists of twenty-five simple examples and problems, and is scored three times the number of correct answers. The synonym-antonym test consists of eighty pairs of stimulus words, forty pairs each of synonyms and antonyms arranged in a chance order by tens. The subject is required to

TABLE 8

Correlations of Thurstone tests with average first year grades, for groups of Engineering and Arts and Sciences students, separately and combined

	THURSTONE IV				THURSTONE, 1920 EDITION			
	Number of cases	Average	S.D.	r	Number of cases	Average	S.D.	r
Arts and Sciences.	107				91			
Grades.....		7.3	0.9			7.4	0.9	
Total score.....		87.8	20.6	0.261 ± 0.06		84.2	23.1	0.387 ± 0.06
Engineers.....	86				59			
Grades.....		7.5	0.9			7.5	0.9	
Total score.....		85.8	23.3	0.372 ± 0.06		82.4	23.7	0.343 ± 0.07
Combined group..	193				150			
Grades.....		7.4	0.9			7.4	0.9	
Total score.....		86.9	21.9	0.317 ± 0.04		83.9	23.3	0.366 ± 0.04

indicate whether the two words of a pair are similar or opposite in meaning by writing the letter 's' or 'o' on the line between the two words of a pair. The right minus wrong method is used in scoring. A completion test of a somewhat different form is the third test in the series. The number of elisions varies from two to six for different sentences, and the subject is given a choice of one of four words which will complete adequately each elision in a sentence. The score is the number of correct selections. The fourth test is an information test of the true-false type, including 40 items, an examination of which shows that the information

tested is drawn from American and European history, geography, English literature, technical subjects such as physics, chemistry, and biology.

In contrast to the Thurstone information tests, approximately 30 of the 40 items are identical with or very similar to items included in the Iowa high school content examination (22) which is based directly on high school curricula. Thus a smaller proportion of items is based on the information acquired incidentally or outside of formal school work. It is distinctly not a vocabulary test of the traditional sort in that only seven to ten of the forty items would probably ever be included in a usual vocabulary test.

At least two of the four tests are among those conceded to be language tests, *i.e.*, the synonym-antonym and completion tests, in both of which it would seem that a knowledge of the meaning of the stimulus words is a primary factor conditioning success. In the synonym-antonym test, the problem that faces the subject is not merely making a judgment of the relationship between two familiar words, but rather, understanding the meaning of seemingly difficult words thoroughly enough to be able to judge their relation to each other without any context as a guide. In the completion test, on the other hand, there is a smaller proportion of abstract words and the ability to understand language is probably less of a conditioning factor than in the synonym-antonym test.

Analysis of the interrelations of the sub-tests in the Anderson Psychological Examination

Data on Form II of this examination have been obtained for a group of 210 entering students, examined in October, 1922; and on Forms I and II for a group of 143 freshmen, members of an English composition class, tested in February, 1922, after 21 of the poorest students had been dropped from the university because of failures. The second group is, therefore, selected both as to the course of study pursued, since it includes few engineering students, and through the dropping of the poorest students. For these reasons, a study of the inter-relations of the four sub-tests will be based on the first group, *i.e.*, the entering students.

The range, average, and standard deviation of the total score and of the score in each of the four sub-tests are given in table 9. Coefficients of correlation for each of the sub-tests with the total score and with each other test are given in table 10. The rank order of the coefficients of correlation with the total score, from the highest to the lowest is as follows: synonym-antonym (0.867),

TABLE 9

Range, average and standard deviation of the total score and score in each of the four tests in the Anderson Psychological Examination, Form II, for a group of 210 entering students

	RANGE	AVERAGE	S.D.	POSSIBLE SCORE
Total score	39 to 154	97.3	24.4	235
Test 1, arithmetic	12 to 63	38.5	8.6	75
Test 2, synonym-antonym	−2 to 60	28.3	12.6	80
Test 3, completion	3 to 37	19.1	5.9	40
Test 4, information	−6 to 30	11.3	6.7	40

TABLE 10

Coefficients of correlation between total score and sub-tests in the Anderson Psychological Examination, Form II, for a group of 210 entering students

	TOTAL SCORE	TEST 1 ARITHMETIC	TEST 2 SYNONYM-ANTONYM	TEST 3 COMPLETION	TEST 4 INFORMATION
Total score		0.592±0.03	0.867±0.01	0.697±0.02	0.608±0.02
Test 1, arithmetic	0.592±0.03		0.248±0.04	0.352±0.04	0.078±0.04
Test 2, synonym-antonym	0.867±0.01	0.248±0.04		0.534±0.03	0.411±0.03
Test 3, completion	0.697±0.02	0.352±0.04	0.534±0.03		0.207±0.04
Test 4, information	0.608±0.02	0.078±0.04	0.411±0.03	0.207±0.04	

completion (0.697), information (0.608), arithmetic (0.592). That is, the synonym-antonym test has most and the arithmetic test least in common with the other tests in the series.

Of the inter-test correlations the highest is that between the synonym-antonym and completion tests (0.534) and the lowest, that between the arithmetic and information tests (0.078). It

would seem that there is some justification for the *a priori* assumption that the synonym-antonym and completion tests have more in common with each other than with the arithmetic and information tests. Regardless of whatever else they may have in common, the ability to understand rather difficult words is required in both tests. If we knew the extent to which the synonym-antonym test is a measure of vocabulary, we could more easily explain the fact that the information test correlates more highly with synonym-antonym than with either of the other tests. It

TABLE 11

Coefficients of partial correlation between the tests in the Anderson Psychological Examination, Form II, for a group of 210 entering students

Arithmetic = 1; synonym-antonym = 2; completion = 3; information = 4.

	CORRELATION COEFFICIENTS (ZERO ORDER)		CORRELATION COEFFICIENTS (FIRST ORDER)
12	0.248±0.04	12.3	0.075±0.04
13	0.352±0.04	12.4	0.237±0.04
14	0.078±0.04	13.2	0.268±0.04
23	0.534±0.03	13.4	0.344±0.04
24	0.411±0.03	14.2	−0.027±0.04
34	0.207±0.04	14.3	0.005±0.04
		23.1	0.493±0.03
		23.4	0.500±0.03
		24.1	0.405±0.03
		24.3	0.363±0.04
		34.1	0.192±0.04
		34.2	−0.016±0.04

has been said that the vocabulary and information tests are related in that they both involve the ability to discover and remember the meanings of words, names, events, etc. It seems probable that this ability accounts in part for the correlation between the two tests under discussion.

An examination of the partial correlation coefficients in table 11 gives further evidence as to the relations of these tests. The correlation of the synonym-antonym and completion tests is but slightly reduced by the removal of the factor common to the information test and only slightly further reduced by the removal

of factors common to the two tests and arithmetic. There is a larger common factor between the three verbal tests, synonym-antonym, completion and information, than between the synonym-antonym, completion and arithmetic reasoning tests.

Considering the interrelations of the synonym-antonym, completion, and information tests, if the effect of the information test is removed, the coefficient of correlation between the remaining two tests is 0.500; if the effect of the completion test is removed the coefficient is 0.363; if the effect of the synonym-antonym test is removed, the coefficient is −0.016. That is, the correlation existing between these three tests is largely due to some factor inherent in the synonym-antonym test.

If a similar analysis of the interrelations of the other possible combinations of tests is made it becomes apparent that, in general, factors common to the synonym-antonym test contribute most, and those common to the arithmetic test least to the inter-correlations of the four tests. Or, in other words, the total test score is heavily weighted with factors inherent in the synonym-antonym test.

Correlations with grades. In order to obtain groups large enough to warrant the use of the correlation method, all of the freshmen for whom there are records in the Anderson examination, including the entering students and the English composition group have been grouped together for the correlations of total score and scores in sub-tests with grades in particular college courses, but they have not been grouped together for correlations with average grades, for reasons that will become apparent. It will be remembered that the entering students were tested with Form II, whereas the English composition group was given both forms. The two forms are presumably equivalent and it was concluded that it would be better to use the scores in Form I for the English composition group, rather than those in Form II which was given after Form I and which may show practice effects.

Correlations have been derived for the total test score, and for each sub-test with first semester grades in English composition, mathematics, physics, chemistry, biology, French and history. The range, average, and standard deviation for each test and the

total score, for each group in each subject are given in table 12; those for the subjects in table 13; and the correlation coefficients in table 14. The coefficients are not sufficiently high to warrant hard and fast conclusions, but they indicate definite tendencies.

TABLE 12

Range, average and standard deviation of total score and of scores in each of the sub-tests of the Anderson Psychological Examination, in each subject group

	ENGLISH COMPOSITION 271 CASES	MATHEMATICS 207 CASES	PHYSICS 171 CASES	CHEMISTRY 147 CASES	BIOLOGY 96 CASES	FRENCH 151 CASES	HISTORY 127 CASES
Total score:							
Range	39–174	39–174	49–174	39–174	61–158	51–174	42–174
Average	103.0	102.0	104.5	103.6	106.5	104.5	105.3
S.D.	25.0	24.5	25.3	24.9	24.1	23.8	24.5
I. Arithmetic:							
Range	12–63	12–63	21–63	18–63	18–60	15–63	12–63
Average	40.8	41.1	42.4	42.3	40.9	40.7	40.1
S.D.	8.6	8.4	9.0	8.7	9.2	8.8	8.4
II. Synonym-antonym:							
Range	4–60	4–60	−2–60	−1–59	9–58	7–59	10–60
Average	30.5	30.0	29.8	29.2	32.6	31.7	32.4
S.D.	12.3	10.4	14.0	13.5	12.0	11.9	12.0
III. Completion:							
Range	4–37	4–37	3–37	4–37	7–37	7–37	7–34
Average	20.4	20.1	20.6	20.4	20.9	20.3	20.2
S.D.	5.9	5.9	6.0	6.2	6.1	5.9	5.4
IV. Information:							
Range	−6–30	−6–30	−6–28	−6–28	−6–28	−2–30	−2–30
Average	11.8	11.3	12.0	11.7	12.0	12.2	12.8
S.D.	6.8	6.4	6.6	6.6	6.7	6.3	7.0

The total test score correlates more highly with grades in English composition (0.485) than with any other subject, and less well with chemistry (0.216) and biology (0.237) The arithmetic test correlates most highly with mathematics (0.377), physics (0.365) and chemistry (0.342) grades. The synonym-

antonym test correlates most highly with English (0.403), history (0.409), and French (0.323) grades. Sentence completion correlates only slightly better with English grades (0.277) than with mathematics (0.259) and physics (0.239). Finally, information correlates most highly with English composition (0.304) and history (0.207).

TABLE 13

Range, average, and standard deviation of first semester grades of college freshmen tested with the Anderson Psychological Examination

SUBJECT	NUMBER OF CASES	RANGE	AVERAGE	S.D
English Composition	271	5–10	7.37	1.17
Mathematics	207	2–10	7.24	1.57
Physics	171	5– 9.5	6.91	1.21
Chemistry	147	4–10	7.87	1.28
Biology	96	5– 9	7.08	0.77
French	151	3–10	7.46	1.43
History	127	4–10	6.97	1.40

TABLE 14

Coefficients of correlation between total test score and scores in sub-tests of the Anderson Psychological Examination and first semester grades in college subjects

SUBJECT	NUMBER OF CASES	TOTAL SCORE	I ARITHMETIC	II SYNONYM-ANTONYM	III COMPLETION	IV INFORMATION
English Composition	271	0.485±0.03	0.337±0.03	0.403±0.03	0.277±0.03	0.304±0.03
Mathematics	207	0.322±0.04	0.377±0.04	0.198±0.04	0.259±0.04	0.126±0.04
Physics	171	0.327±0.04	0.365±0.04	0.295±0.04	0.239±0.04	0.155±0.05
Chemistry	147	0.216±0.05	0.342±0.04	0.113±0.05	0.050±0.05	0.140±0.05
Biology	96	0.237±0.06	0.195±0.06	0.245±0.06	0.165±0.06	0.193±0.06
French	151	0.304±0.04	0.271±0.05	0.323±0.04	0.192±0.05	0.051±0.05
History	127	0.386±0.05	0.210±0.05	0.409±0.04	0.143±0.05	0.207±0.05

If the coefficients of correlation for total score and for each test with each of the subjects are averaged, the rank order is as follows: 1, total score (0.326); 2, arithmetic (0.299); 3, synonym-antonym (0.283); 4, completion (0.189); 5, information (0.168).

If the coefficients of correlation for each of the four tests with grades in college subjects are averaged separately for each subject, the rank order of the averages differs somewhat from the rank order of the correlations between total test score and grades in subjects. This rank order which indicates the predictive value of the combinaton of tests for ability in specific subjects is as follows: 1, English composition (0.330); 2, physics (0.263); 3, history (0.242); 4, mathematics (0.240); 5, French (0.209); 6, biology (0.199); 7, chemistry (0.161).

An examination of the coefficients reported shows that the subjects tend to fall into two groups: (*a*) English, French, history and biology; (*b*) mathematics, physics, and chemistry. The first group is made up of subjects which involve no mathematics;

TABLE 15

Coefficients of correlation between the sub-tests of the Anderson Psychological Examination and grades in college subjects averaged for two groups of subjects

	GROUP I—ENGLISH, FRENCH, HISTORY, BIOLOGY		GROUP II—MATHEMATICS, PHYSICS, CHEMISTRY	
	Sub-test	Average *r*	Sub-test	Average *r*
1	Synonym-antonym	0.345	Arithmetic	0.361
2	Arithmetic	0.253	Synonym-antonym	0.202
3	Completion	0.194	Completion	0.182
4	Information	0.189	Information	0.140

the second group includes mathematics and the mathematical sciences. Table 15 shows the coefficients of correlation for the sub-tests averaged for each of the two groups. The synonym-antonym test correlates most highly of the four sub-tests with the English group and second highest with the mathematics group; the arithmetic test correlates most highly with the mathematics group and second highest with the English group. The completion and information tests occupy the same relative position for both groups.

The analysis of the correlations existing between the four tests and grades, considered in relation to the analysis of the interrelations of the tests indicates that the synonym-antonym test is a measure of some ability common to certain groups of subjects,

i.e., it is most predictive of ability in English composition, history and French; less predictive of ability in physics and biology; and least predictive of ability in mathematics and chemistry. If it were agreed that success in the first named subjects, English composition, history and French is more dependent upon "language ability," *i.e.*, a knowledge of the meanings of words and the ability to handle abstract verbal concepts, than is success in the last-named groups, mathematics, physics, biology and chemistry, then it would follow that the synonym-antonym test is a good measure, relatively, of this "language ability." It is more specifically predictive of success in English composition, which is described in the University Register as "a course in prose writing . . . . consisting of a review of usage and the study of the principles of structure and style, in general and as exhibited in particular forms of discourse." The low correlation found to exist between the synonym-antonym and arithmetic tests, 0.248, together with the fact that the former gives an average correlation of only 0.202 with the mathematical group of subjects argues further for the specificity of the synonym-antonym test.

The fact that the information test correlates more highly with the synonym-antonym test than with either of the other tests, together with the fact that it, like the synonym-antonym test correlates more highly with English composition and history than with other subjects, indicates the relationship existing between these two tests.

The sentence completion test, while less predictive of either type of ability than the synonym-antonym test, is relatively more predictive of success in the mathematical group of subjects than is the synonym-antonym test. That is, it is less specifically a measure of language ability, as it is involved in non-mathematical subjects than is the synonym-antonym test. The somewhat similar conclusion of Tolman (56) (that the completion test correlates equally well with English and mathematics) is suggested.

The arithmetic test is not as clearly a measure of mathematical ability as opposed to language ability as the synonym-antonym test is a measure of language ability as opposed to mathematical ability. That is, the arithmetic test correlates more highly with the non-mathematical subjects than the synonym-antonym test correlates with the mathematical group.

On the whole, then, the Anderson test is weighted with sub-tests that make it more predictive of success in non-mathematical subjects such as English composition, history, French, and less predictive of success in such subjects as mathematics, physics and chemistry.

Additional evidence for this conclusion has been derived from a study of the correlation of the average first semester grades for the group of entering students, 1922. The students were grouped according to their registration in the engineering school or in the college of arts and sciences. In the first group, *i.e.*, the engineering students, the course of study is rather fixed and invariable for the first year students including generally, mathematics, engineering drawing, general engineering, English composition, history, and in some cases chemistry and a language, for students entering without advanced credits; and including applied mechanics, mathematics, physics, or chemistry, survey, and in some cases English composition for engineers entering with advanced credit. In the second group, the course of study is more varied including a greater proportion of courses in English, languages, history, and biology, although a large proportion of the group take mathematics. In general, it can be said, then, that the average grades for the engineering group are largely weighted with mathematics and mathematical sciences, while those of the arts and sciences group are weighted with non-mathematical subjects, although the division between the two is not perfect. The averages were derived by the method of weighting according to credits previously described.

Correlation between test scores and grades for groups of engineering and arts and sciences students separately and combined

Correlations of the total Anderson score, each of the sub-tests and combinations of sub-tests with average grades for the engineering and arts and sciences groups separately and together are given in table 16. The total score correlates more highly with the grades of the arts and sciences group than with those of the engineers; the synonym-antonym test alone correlates even more highly with the grades of the arts and sciences group but much less well with those of the engineers; completion also shows a

difference in favor of the arts and sciences group; the arithmetic test is almost equally predictive of both, although the coefficient for the engineers is slightly higher; and the information test shows

TABLE 16

Coefficients of correlation for the total score, sub-tests and combinations of sub-tests in the Anderson Psychological Examination with average grades for the first semester, for groups of Engineers and Arts and Sciences students, separately and combined

	110 ARTS AND SCIENCES STUDENTS			79 ENGINEERS			189 COMBINED		
	Average	S.D.	r	Average	S.D.	r	Average	S.D.	r
Average first semester grades	7.3	0.9		7.1	0.9		7.2	0.9	
Total score	99.7	24.3	0.543 ±0.04	94.6	25.2	0.334 ±0.06	97.6	24.8	0.465 ±0.03
2. synonym-antonym	30.9	10.7	0.559 ±0.04	24.7	13.4	0.182 ±0.07	28.3	12.8	0.374 ±0.04
3. completion	19.6	6.2	0.401 ±0.05	19.0	5.9	0.067 ±0.07	19.3	6.1	0.276 ±0.04
1. arithmetic	38.9	8.1	0.409 ±0.05	41.2	8.7	0.434 ±0.06	39.8	8.4	0.403 ±0.04
4. information	11.7	6.8	0.195 ±0.06	11.1	6.1	0.119 ±0.07	11.4	6.6	0.170 ±0.04
Total score minus test 2, synonym-antonym	69.1	15.4	0.513 ±0.04	69.8	14.9	0.345 ±0.06	69.4	15.2	0.455 ±0.03
Total score minus test 1, arithmetic	62.3	19.2	0.508 ±0.04	54.4	21.6	0.164 ±0.07	59.0	20.6	0.366 ±0.04

an equal lack of correlation with the grades of both groups. If the synonym-antonym test is removed from the total score, while the coefficient of correlation is still higher for the non-

engineers, that for the engineers is relatively slightly higher than the correlation with total score. The removal of the arithmetic test reduces the coefficients for both groups, but that of the engineers relatively more.

In general, the correlations for the total group of subjects, combining both the engineering and arts and sciences students, are about mid-way between those for the two groups separately, which is to be expected since the distributions for the two groups are not very dissimilar. It is obvious, then, that the correlation of the total score and average grades for all students conceals the real nature of correlation. The Anderson test is, apparently, a fairer test for non-engineering students, at least as far as predictability of grades is concerned.

The inclusion of the synonym-antonym test together with the other three tests actually reduces the correlation between total test score and average grades of engineering students. Taken alone it is slightly more predictive of grades of arts and sciences students than the total score. The arithmetic test alone correlates more highly with the average grades of the engineers than does the total test score or the total test minus the synonym-antonym test.

With a division into mathematical and non-mathematical groups as rough as this is admitted to be, we have been able to show that the synonym-antonym test, the completion test to a smaller degree, and the total Anderson test are more predictive of the non-mathematical type of ability than of the mathematical, and correlations discussed above have shown that the ability as involved in English composition work is particularly well predicted. If it were possible to make finer classifications of those individuals taking no mathematics and no science, and those taking only mathematics and sciences, and still have large enough groups to justify the use of the correlation method, the tests might be shown to be more specific still.

It is an inescapable conclusion that an intelligence examination heavily weighted with tests that we have called language tests is not ideally suited to the examination of a group of students including both the students registered in engineering and arts and sciences, if the test score is to be considered as predictive of average

grades, and that there is need of including in intelligence examinations tests material better suited to bring into play abilities specifically inherent in mathematical and scientific subjects, possibly of a difficult mathematical reasoning type.

III. THE JOHNS HOPKINS COMBINATION TEST

The Johns Hopkins Combination test is composed of three parts: 1, a synonym-antonym test made up of the Anderson Form I words, with revised directions; 2, a cancellation test; 3, a group of fourteen tests, including arithmetical reasoning, ingenuity problems, a list of ten hard opposites, and a list of ten difficult words to be defined, chosen from the Terman vocabulary test. Two and a half minutes working time are allowed for part 1 and likewise for part 2. Thirty minutes are allowed for the fourteen tests in part 3, to be distributed among the tests as the subjects wish. The arithmetic reasoning and ingenuity problems are scored right or wrong; opposites are scored one, one-half and zero, which values are based on an evaluation made by six graduate students in psychology and a professor of English, of all of the responses given; the definitions are scored one, one-half, and zero, a value of one being given for a complete definition, one-half for a definition that is correct except for the fact that the part of speech of the stimulus word has been changed, or for an incomplete definition of the correct part of speech, and zero for a wrong definition, or for an incomplete definition together with an incorrect part of speech. All of the definitions have been graded by one person. The possible score in part 3 is 168, 88 for arithmetical reasoning and ingenuity, 40 for opposites, and 40 for definitions. Part 2 has been omitted from the results reported here.

This test combines elements that are distinctly of the abstract verbal type, and elements less distinctly verbal which involve, however, the use of number concepts. In part 3, in which the subjects can solve the problems in any order, it is possible for them to solve one type of problem and omit the other. Therefore, preferences for, or special abilities in specific types of tests should be shown.

Two hundred and twenty-three students entering the Johns Hopkins University in October, 1923, were given the Hopkins

Combination Test. The range, average, and standard deviation for the total score (part 1 and part 3), for part 1, and for separate tests and groups of tests in part 3 are given in table 17. Approximately 60 per cent of the average total score is contributed by the synonym-antonym, opposites and definitions tests, that is, by the distinctly verbal or language parts of the test, and 40 per cent by the arithmetical reasoning and ingenuity problems.

An examination of individual records shows that some of the subjects solved only the arithmetical and ingenuity problems in part 3, while others attempted only the language material. These extreme cases, however, are not numerous. An analysis of the

TABLE 17

Range, average, and standard deviation of total score, score in part 1, groups of tests and single tests in part 3 of the Johns Hopkins Combination Test, for a group of 223 entering students

	RANGE	AVERAGE	S.D.	POSSIBLE SCORE
Total score	24 to 190	113.4	33.8	248
Part 1. Synonym-antonym	−9 to 67	30.8	14.0	80
Part 3. Arithmetical reasoning and ingenuity	0 to 88	44.7	16.0	88
Opposites and definitions	0 to 64	38.6	14.8	80
Opposites	0 to 36	22.3	7.2	40
Definitions	0 to 36	17.5	9.6	40
Synonym-antonym, opposites and definitions	−3 to 121	69.4	25.5	160

responses of each of the 223 individuals in the synonym-antonym, opposites and definitions tests and in arithmetical reasoning and ingenuity as either above or below the average of the group, shows that 19.7 per cent of cases are above the average in all four measures; 16.5 per cent are below in all; 10.7 per cent are above the average in synonym-antonym, opposites and definitions, but below in arithmetical reasoning and ingenuity; 9.4 per cent are below in the three language tests and above in arithmetical ingenuity. Thus 56.3 per cent of the cases fall into definite categories. Of the remainder, 5.8 per cent are above in all of the tests except definitions; 5.4 per cent are below in all except giving opposites; and 4.9 per cent are above in synonym-antonym

and opposites, but below in definitions and arithmetical reasoning and ingenuity; 27.6 per cent are distributed among the other possible combinations.

Thus we see that the largest groups of subjects are either above the average or below in all tests of one type, and the next two largest groups are above the average in all language tests and below the average in the same tests.

Correlations between these four tests and combinations of tests which express the relationship more concisely, although not with the same fullness of meaning as the per cents give, are shown in table 18. The degree of correlation between any two of the verbal or language tests is practically equal (0.522 to 0.562) although it is not sufficiently high for one to be substituted for

TABLE 18

Intercorrelations of sub-tests in the Johns Hopkins Combination Test for a group of 223 entering students

	SYNONYM-ANTONYM	OPPOSITES AND DEFINITIONS	SYNONYM-ANTONYM, OPPOSITES AND DEFINITIONS	OPPOSITES
Opposites	0.556±0.03			
Definitions	0.562±0.03			0.522±0.03
Opposites and definitions	0.545±0.03			
Arithmetical reasoning and ingenuity	0.247±0.04	0.234±0.04	0.248±0.04	

the other. The partial coefficients of correlation between synonym-antonym and opposites with definitions held constant is 0.372; between synonym-antonym and definitions with opposites held constant, 0.383; between definitions and opposites with synonym-antonym held constant, 0.304. Thus the removal of the effect of one of the tests from the correlation of the other two reduces the degree of correlation to about the same extent in all three combinations of the tests.

The degree of correlation between the language tests, singly or together, with arithmetical reasoning and ingenuity is practically the same (0.234 to 0.248) and considerably lower than the intercorrelation of the language tests themselves. There is a distinct difference, therefore, between factors measured by the two types of test material.

Correlations between the two types of test and grades in English composition and mathematics (table 19) show that the synonym-antonym test alone, and opposites and definitions together, are more predictive of grades in English composition than in mathematics, which has been true for all other coefficients of correlation between similar tests and grades. The combination of the synonym-antonym with opposites and definitions tests increases the correlation with grades in English composition, and to a slight degree, the correlation with mathematics grades.

Arithmetical reasoning and ingenuity tests correlate more highly with mathematics than with English composition grades, but

TABLE 19

Coefficients of correlation of first semester grades in English composition and mathematics and with sub-tests and combinations of sub-tests of the Hopkins Combination Test for entering students

TEST	ENGLISH COMPOSITION, 142 CASES			MATHEMATICS, 127 CASES		
	Average	S.D.	r	Average	S.D.	r
Synonym-antonym	31.8	12.7	0.484±0.04	32.3	12.5	0.374±0.05
Opposites and definitions	40.0	14.7	0.443±0.04	39.6	14.0	0.370±0.05
Synonym-antonym, opposites and definitions	70.9	25.4	0.531±0.04	71.3	24.4	0.397±0.05
Arithmetical reasoning and ingenuity	45.0	16.2	0.143±0.05	44.9	14.9	0.332±0.05

less well with mathematics than does the synonym-antonym test or the combination of language tests. The reason for this is not altogether clear. It may be associated with the scoring of the items of the arithmetical reasoning and ingenuity type, and the weighting assigned to the different items, although there is no experimental evidence, as yet, for this assumption.

Correlations between tests and average grades for groups of engineers and arts and sciences students separately and together

Correlations of average first semester grades with the total test score, scores in sub-tests and groups of sub-tests have been

derived for groups of engineering and arts and sciences students, separately and together, similar to those for the groups tested with the Thurstone and Anderson examinations (table 20).

Similarly to the Anderson examination, the Hopkins examination is more predictive of the grades of arts and sciences than of engineering students, although there is less difference between

TABLE 20

Coefficients of correlation for the total score, sub-tests, and combinations of sub-tests in the Johns Hopkins Combination Test with average first semester grades, for groups of engineering and arts and sciences freshmen, separately and together

	130 ARTS AND SCIENCES STUDENTS			80 ENGINEERS			210 TOTAL GROUP		
	Average	S.D.	r	Average	S.D.	r	Average	S.D.	r
Average first semester grades	7.3	0.9		7.3	0.9		7.3	0.9	
Total test score	113.4	34.4	0.501 ±0.03	116.1	33.4	0.468 ±0.05	114.6	34.0	0.482 ±0.03
Synonym-antonym	31.8	12.9	0.515 ±0.04	29.8	14.0	0.394 ±0.06	31.1	13.4	0.438 ±0.03
Synonym-antonym, opposites, and definitions	71.1	25.7	0.514 ±0.04	65.2	25.1	0.378 ±0.06	68.9	25.6	0.464 ±0.03
Arithmetical reasoning and ingenuity	42.6	16.1	0.257 ±0.05	49.8	14.3	0.263 ±0.07	45.4	15.8	0.245 ±0.04

the coefficients for the groups tested with the Hopkins examination than those tested with the Anderson examination. The Hopkins examination stands between the Anderson examination, which differentiates between the two groups clearly, and the Thurstone 1920 edition, which differentiates between the groups still less than the Hopkins test does. The Thurstone IV Examination is the only one of the four that shows a higher coefficient with grades

for the engineers than for the arts and sciences group, and again, the difference is not marked.

As was true of the synonym-antonym test in the Anderson examination, so in the Hopkins test group, it is more predictive of grades of the arts and sciences group than of the engineering group. The arithmetical reasoning and ingenuity problems in the Hopkins test are almost equally predictive of the grades of both groups, although the coefficient of correlation with grades is slightly higher for the engineers, as is the arithmetic test in the Anderson examination. But the combination of tests in the Hopkins examination is such that the total score does not favor the arts and sciences group to the same extent that the Anderson examination does.

It was found for the groups tested with the Anderson synonym-antonym test that the presence of the synonym-antonym test score in the total score actually lowered the correlation with average grades of engineering students. That the synonym-antonym test is not a part of the Thurstone examination may account in part for the relatively higher degree of correlation of this examination with average grades of engineering students. That the Thurstone examination is relatively less predictive of grades of the arts and sciences group than are the Anderson and Hopkins examinations, may be due to the fact that it includes a larger proportion of test material which, although verbal in nature, involves reasoning as in the analogies, syllogisms and reading tests to a greater extent than it is involved in the language tests of the Anderson and Hopkins examinations.

This discussion suggests inevitably the necessity of studying the type of test material included in psychological examinations in relation to the special abilities or interests of the subjects to be tested, at least as represented by their grades in college subjects, inasmuch as we have been able to give specific evidence that some tests are more predictive of a type of ability involved in a particular subject than in other subjects. At the present time, specific tests for predicting the ability involved in such subjects as English composition, *i.e.*, language ability of some sort, are better measures than those tests that predict success in mathematics and the mathematical sciences.

An apparent exception to this statement is found in Wood's study of the Thorndike Intelligence Examination (69) in which he has shown that a group of mathematical tests including arithmetic fundamentals and reasoning, number completion, and algebra, requiring 26 minutes, gives a correlation coefficient with mathematics grades of 0.66 for students who have completed two years of college work in Columbia University; whereas a group of language tests including directions, disarranged sentences, opposites and absurdities, requiring 16 minutes, gives a correlation coefficient of only 0.42 with English grades. He has shown, however, that the mathematical tests give an excellent distribution, whereas the language tests give a decidedly skewed distribution, seeming to indicate that they are too simple to differentiate between college students; and that the coefficient of reliability for mathematics grades (equal to the correlation coefficient between grades for two semesters) is higher (0.61) than that for English grades (0.36). This fact is significant inasmuch as Kelley (28) has shown that the reliability coefficients set a limit upon the correlation between two variables, since, except in spurious chance cases, the maximum possible correlation between two variables is "the product of the square roots of their reliability coefficients." Thus the low degree of correlation between the language tests and English grades is probably due to the combined effects of the poor distribution of the scores in the language tests and the low reliability coefficient of English grades. The Thorndike examination is suggestive of the type of material that is needed for predicting ability in mathematics.

CHAPTER III

Synonym-Antonym Tests at the Johns Hopkins University

Inasmuch as the synonym-antonym test has given evidence, both in this study and in other investigations, of being a superior measure particularly of ability involved in such college work as English Composition, and likewise of general intelligence, especially of high school pupils, but of college students as well, a detailed analysis of the test has been deemed both justified and desirable.

I. DESCRIPTION OF TEST AND OF GROUPS TESTED

Records have been obtained for several groups of college students in both forms of the synonym-antonym test arranged by Anderson, and included in his psychological examination.

The form of test used by Anderson is similar to that in the National Intelligence test, Form A, and differs from the Army Alpha form in that the subject is directed to write the letter "S" on the line between the two words of a pair if they are similar in meaning, or the letter "O" if they are opposite in meaning, instead of underlining the words "similar" or "opposite" as in Alpha. Each form consists of 80 pairs of words, 40 pairs each of synonyms and antonyms, arranged in a chance order in groups of ten. Although the question of the relative difficulty of the stimulus words will be discussed in more detail somewhat later, it may be said at this point that Anderson has omitted from his lists of words the easiest pairs of words listed in the Alpha forms. He has drawn heavily from Alpha for his stimulus words, but has included some words that do not appear in Alpha. The data on the groups tested do not indicate that the pairs of words have been arranged in any order of difficulty. The time allowed is two and one-half minutes.

The right minus wrong method has been used in scoring. In the type of test in which there are only two possible responses,

the assumption is that a subject not knowing any of the responses will get half of them right by chance; or of the total number of items which the subject does not know but checks at random, one-half will be correctly and one-half incorrectly checked. Thus, one of the correct responses is subtracted for each of the incorrect responses in order to give the best measure of the actual knowledge of the subject. A criticism of this method of scoring will be offered later but for the present it will be used.

Records of performance in the synonym-antonym test have been obtained for the following groups under the conditions stated: (*a*) 143 freshmen, members of an English composition class were given Forms I and II of the Anderson Psychological Examination, during one class period in February, 1922, after the poorest students had been dropped from the University because of failures. The group, therefore, represents a selection in addition to its selection as an English composition class. It will be referred to as the English Composition Group. (*b*) Forty-seven members of an introductory psychology class, made up of students in all four college classes, but chiefly of upper classmen, were tested with Form I and the first three tests in Form II of the same psychological examination, during one class period, in February, 1922. They will be designated Psych. 1, 1922. (*c*) As part of the psychological examination required of entering students, the Anderson Psychological Examination, Form II, was given to 210 entering students in October, 1922. (*d*) Forty-three members of an introductory psychology class were tested with 320 pairs of synonyms and antonyms, including the 160 pairs of words in the two Anderson forms, but not in the order as used by Anderson. The time allowance was not altogether comparable to that of the other groups. (*e*) Two hundred and twenty-three entering students were tested in October, 1923, with a combination of tests, the first of which was Anderson's Form I list of synonyms and antonyms, with some slight changes in directions. Data have been obtained, therefore, for a total of 666 cases.

These last two groups will be discussed in more detail in connection with the question of the effects of retests and changing the form of the make-up and the administration of the test, and will therefore be mentioned only incidentally in this connection.

The change in directions used with the last group was not deemed sufficient to render invalid a comparison of the results of this group with those of the other groups.

II. DISTRIBUTIONS OF SCORES AND OF RESPONSES TO INDIVIDUAL PAIRS OF STIMULUS WORDS

Histograms showing the distribution of scores in Forms I and II are presented in figures 1 to 6, and decile distributions in table 21. For the two groups which were given both forms of the test on the same day, *i.e.*, Eng. Comp. and Psych. 1, 1922, there is an increase in average score in Form II, and a higher score at every

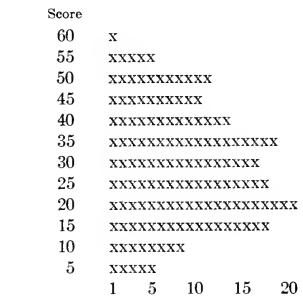

FIG. 1. DISTRIBUTION OF SCORES IN THE ANDERSON SYNONYN-ANTONYM TEST, FORM I, FOR ENGLISH COMPOSITION GROUP

decile, except the first decile for Eng. Comp., which indicates either that Form II is less difficult than Form I or, more probably, that there is a practice effect. Since none of the pairs of words are repeated, the practice effect if there is one, must be in adjustment to the conditions of the test and in the method of making the necessary judgment as to the relation existing between the pairs of words.

The decile distributions (table 21) of the three groups tested with Form I are very similar. It is interesting to note the small difference between the average performance of the three groups which represent entering students, students who have had a half year of college training, and upper classmen.

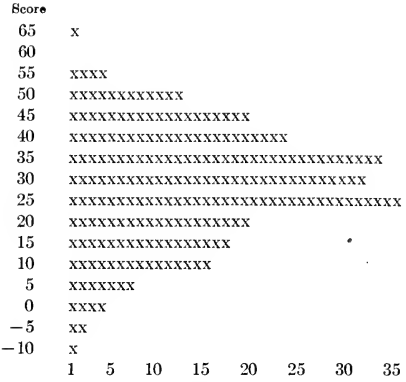

FIG. 2. DISTRIBUTION OF SCORES IN THE ANDERSON SYNONYM-ANTONYM TEST, FORM I, FOR ENTERING STUDENTS, 1923

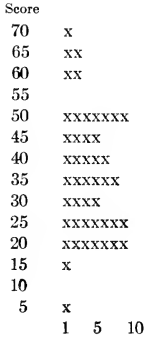

FIG. 3. DISTRIBUTION OF SCORES IN THE ANDERSON SYNONYM-ANTONYM TEST, FORM II, FOR PSYCHOLOGY I, 1922 GROUP

The distributions of scores of the groups tested with Form II are not as similar, which is to be expected since two of the three groups have had practice in Form I. The lowest scores at each

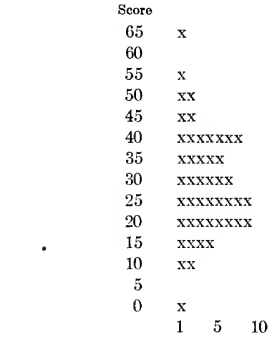

FIG. 4. DISTRIBUTION OF SCORES IN THE ANDERSON SYNONYM-ANTONYM TEST, FORM I, FOR PSYCHOLOGY I, 1922 GROUP

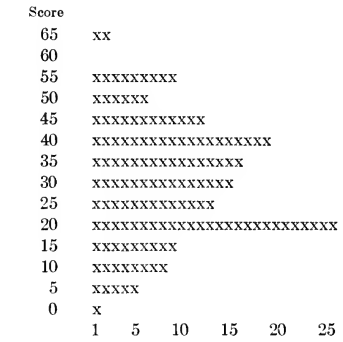

FIG. 5. DISTRIBUTION OF SCORES IN THE ANDERSON SYNONYM-ANTONYM TEST, FORM II, FOR ENGLISH COMPOSITION GROUP

decile are those of the entering students of 1922, for Form II, who had had no practice.

The coefficient of reliability for the Eng. Comp. group is 0.792 which indicates that whereas the two forms yield fairly similar scores, they are not exactly equivalent.

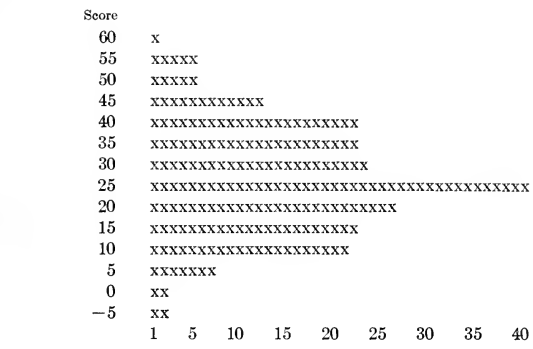

FIG. 6. DISTRIBUTION OF SCORES IN THE ANDERSON SYNONYM-ANTONYM TEST, FORM II, FOR ENTERING STUDENTS, 1922

TABLE 21

Decile distribution of scores in the Anderson synonym-antonym test

	FORM I			FORM II		
	Eng. comp.	Psych. 1, 1922	Entering students, 1923	Eng. comp.	Entering students, 1922	Psych. 1, 1922
Number of cases	141	47	223	141	210	47
Upper limit	60.0	66	67	67	60	70
9th decile	51.3	48.2	48.6	52.4	45.8	60.7
8th decile	44.5	42.5	43.5	45.3	40.6	51.8
7th decile	39.3	38.9	38.8	41.5	35.9	47.3
6th decile	35.4	34.3	35.4	37.8	31.3	42.2
5th decile	31.0	30.4	31.8	32.8	28.1	37.9
4th decile	26.5	27.3	28.4	27.8	25.5	33.5
3rd decile	23.0	24.8	20.5	23.7	21.7	28.6
2nd decile	19.4	21.5	19.5	21.0	17.2	25.2
1st decile	15.3	17.1	12.7	10.0	12.3	21.9
Lower limit	5.0	7.0	−9.0	3.0	−2.0	6.0

A distribution of the responses to each of the pairs of stimulus words has been made for all of the groups tested in order to determine in part the relative difficulty of the individual synonyms and antonyms, and whether it remains constant for groups

differing in training and selection; and to discover any characteristics of response. It is recognized that in the type of test in which there are only two possible responses, the chance factor enters in, but it is assumed that it plays an equal part in the responses to all of the stimulus words, and can therefore be disregarded in a discussion of the relative difficulty of the pairs of words. It was found that the time limit was so short that the number of cases responding to each pair of words decreased very rapidly after the first forty pairs of words, and considerably less than half of the members of each group responded to more than fifty pairs.

The percentages of accuracy (the per cent correct of the total number of items attempted) have been determined for each pair of words in Forms I and II for each of the three groups separately. They are recorded for the first forty pairs of words in the lists, with the number of cases on which the percentages are based in tables 22 and 23. For the Eng. Comp. group, the percentages range from 65 to 100 per cent. The percentage of accuracy for 17 of the 40 pairs is 90 per cent or more; for 32 of the 40 pairs, it is 80 per cent or more. In the Psych. 1, 1922 group, the range of percentages is from 68 to 100 per cent. The percentage of accuracy for 20 of the 40 pairs of words is 90 per cent or more; for 33 of the 40 pairs it is 80 per cent or more. The range of percentages of accuracy in the entering group is from 66 to 96 per cent. The percentage of accuracy of 12 of the 40 pairs of words is 90 per cent or more, and for 28 of the 40 pairs it is 80 per cent or more.

The range of percentage of accuracy for the pairs of words in Form II for the 1922 entering students, is from 58 to 96 per cent. The percentage of accuracy for 11 of the 40 pairs of words is 90 per cent or above, and for 26 of the forty pairs it is 80 per cent or above. The range of percents for the remaining two groups will not be discussed in detail since they represent the second test of the same type and are therefore not directly comparable to the other groups which represent the first test.

It has already been stated that chance will account for some of the correct responses and, therefore, it is not surprising that the percents of accuracy do not fall lower than they do, but it would

TABLE 22

Percentages of accuracy of the first 40 pairs of synonyms and antonyms in Form I, Anderson Examination

STIMULUS WORDS	ENG. COMP. GROUP		PSYCH. I, 1922 GROUP		ENTERING STUDENTS, 1923	
	Number of cases	Per cent	Number of cases	Per cent	Number of cases	Per cent
1. Masculine-feminine	143	100	47	96	223	96
2. Mastery-ascendancy	143	70	47	83	223	73
3. Vanity-conceit	143	82	47	83	223	80
4. Acquire-lose	143	99	47	100	223	98
5. Default-defalcation	143	83	47	91	223	79
6. Decadence-decline	143	88	47	96	223	86
7. Anxiety-nonchalance	143	82	47	94	223	88
8. Merriment-gloom	143	97	47	94	223	98
9. Knave-villain	143	89	47	91	223	86
10. Accountable-irresponsible	143	92	47	94	223	90
11. Greediness-cupidity	143	72	47	68	223	66
12. Immune-susceptible	143	87	47	89	223	82
13. Adapt-conform	143	92	47	98	223	85
14. Pompous-ostentatious	143	83	47	92	223	71
15. Disgust-aversion	143	65	47	72	223	71
16. Momentous-immaterial	143	88	47	98	223	81
17. Specific-general	143	90	47	94	223	85
18. Waste-conserve	143	97	47	98	223	97
19. Any-none	143	87	47	87	223	90
20. Desultory-rambling	143	77	47	87	222	72
21. Avert-prevent	142	90	47	89	221	87
22. Defile-purify	142	94	47	94	219	92
23. Ghost-substance	140	91	46	93	218	89
24. Vestige-trace	140	86	46	85	216	81
25. Acme-climax	139	74	46	87	215	79
26. Accumulate-dissipate	136	95	45	100	214	94
27. Obedient-refractory	135	83	44	79	213	85
28. Apprehensive-fearful	134	73	42	79	209	79
29. Moderate-violent	131	98	41	98	206	96
30. Traditional-legendary	130	93	40	98	206	92
31. Hoax-deception	128	90	39	77	200	90
32. Hasty-circumspect	125	84	37	68	197	71
33. Blot out-delete	124	91	35	89	192	85
34. Irregular-eccentric	123	79	35	80	190	86
35. Amiable-surly	120	92	34	94	189	93
36. Suave-brusque	111	84	32	87	182	80
37. Superfluous-essential	110	94	31	97	177	88
38. Apathy-indifference	106	71	30	83	166	70
39. Pertinacious-obstinate	95	84	29	86	158	77
40. Vilify-praise	89	82	29	90	152	72

TABLE 23

Percentages of accuracy of the first 40 pairs of synonyms and antonyms in Form II, Anderson Examination

STIMULUS WORDS	ENG. COMP. GROUP		PSYCH. 1, 1922 GROUP		ENTERING STUDENTS, 1923	
	Number of cases	Per cent	Number of cases	Per cent	Number of cases	Per cent
1. Often-seldom	143	97	47	98	210	96
2. Largess-donation	143	73	47	83	210	71
3. Degradation-humiliation	143	92	47	89	210	83
4. Con-pro	143	94	47	91	210	89
5. Indict-arraign	143	77	47	81	210	80
6. Radiance-effulgence	143	80	47	85	210	78
7. Depressed-elated	143	92	47	98	210	92
8. Vagrant-dweller	143	93	47	94	210	83
9. Orifice-aperture	143	78	47	91	210	76
10. Transient-permanent	143	92	47	96	210	90
11. Linger-loiter	143	97	47	96	210	96
12. Irksome-refreshing	143	97	47	98	210	98
13. Recoup-recover	143	93	47	87	210	90
14. Confer-grant	143	90	47	89	210	85
15. Plenary-complete	143	66	47	70	210	58
16. Actual-fictitious	143	99	47	98	210	95
17. Facility-difficulty	143	98	47	98	210	95
18. Integrity-dishonesty	143	90	47	94	210	84
19. Besmirch-cleanse	143	94	47	85	210	90
20. Obdurate-stubborn	143	89	47	85	210	89
21. Lucrative-profitable	143	76	47	85	206	77
22. Significant-paltry	143	85	47	91	206	78
23. Discretion-folly	143	89	47	89	206	86
24. Tedious-dry	143	80	47	85	206	74
25. Torpor-stupor	143	82	47	87	205	76
26. Imprisoned-free	141	98	47	100	204	99
27. Carnivorous-herbivorous	139	79	47	77	198	80
28. Influence-incentive	139	80	47	81	196	76
29. Aggrandize-belittle	137	74	46	89	190	64
30. Facilitate-further	136	79	46	93	189	85
31. Scarcity-dearth	134	75	46	74	187	76
32. Insipidity-zest	134	79	46	89	183	80
33. Null-void	134	87	45	91	182	86
34. Avarice-cupidity	133	74	45	84	175	65
35. Sterile-fertile	132	73	44	86	171	79
36. Terrestrial-celestial	129	89	43	88	164	85
37. Convoke-dismiss	125	75	41	76	159	61
38. Mysterious-cryptic	125	75	40	78	150	73
39. Latent-hidden	120	78	39	87	143	84
40. Straight-tortuous	116	93	39	90	136	93

also seem that a large percentage of the stimulus words are fairly easy. It would be impossible to rank the stimulus words in a definite order of difficulty without measuring the time of each individual response separately.

In order to determine whether the same pairs of words offer equal difficulty to different groups of individuals, coefficients of correlation between the percentages of accuracy for the first 40 pairs of words in the different groups tested were derived. The percents of accuracy were derived for the Psych. 1, 1923 group, and the correlation with other groups reported although it is admitted that this group is not strictly comparable to the other groups. The coefficients are as follows:

1. Eng. Comp. 1 with entering students 1923, Form I..... 0.845
2. Eng. Comp. 1 with entering students 1922, Form II..... 0.908
3. Eng. Comp. 1 with Psych. 1, 1922, Form I..... 0.708
4. Entering students 1922 with Psych. 1, 1923, Form II.... 0.718
5. Psych. 1, 1922, with Psych. 1, 1923, Form I..... 0.679
6. Psych. 1, 1922, with Psych. 1, 1923, Form II..... 0.642

The correlation is greatest between the freshmen groups which are the largest and lowest between the Psych. 1 groups, and intermediate between the two extremes for the freshmen groups with the Psych. 1 groups. These coefficients suggest that the pairs of synonyms and antonyms are of the same relative difficulty for large groups of college freshmen. The considerably smaller degree of correlation between the Psych. 1 groups may indicate that upper-classmen are specialized along different lines and the relative difficulty of the words is not then identical for the two groups.

If the percents of accuracy of the first 40 pairs of words are averaged separately for synonyms and antonyms, the antonyms show a higher average in each group (table 24). The significance of this difference is discussed in a succeeding section.

III. ANALYSIS OF THE PERFORMANCE OF A SMALL GROUP IN A RETEST OF THE SYNONYM-ANTONYM TEST

Four lists of 80 stimulus words each, 40 synonyms and 40 antonyms were prepared, each list including 40 pairs of words from

Anderson's lists, together with 40 additional pairs. The form of the test was the same as that used by Anderson, including the directions, which were as follows:

If the two words of a pair are similar in meaning, that is, may be substituted for one another in a sentence without seriously altering the meaning, write the letter "S" on the line between the two words. If they are opposite in meaning, write the letter "O" on the line between the two words. If you cannot be sure, guess. The samples are already marked as they should be.

Samples: good O bad
little S small

TABLE 21

Average percentages of accuracy for the first forty pairs of antonyms and synonyms in Forms I and II, of the Anderson Psychological Examination

GROUP	FORM I			FORM II		
	Number of cases	Average per cent of accuracy		Number of cases	Average per cent of accuracy	
		20 antonyms	20 synonyms		20 antonyms	20 synonyms
English Composition	143	90.8	81.6	143	89.0	81.0
Psychology, 1922	47	92.2	85.7	47	91.2	85.1
Psychology, 1923	43	95.2	90.7	43	94.4	85.9
Entering students, 1923	223	88.3	79.8			
Entering students, 1922				210	85.8	78.9
Total				435	88.5	80.9

These four lists, making a total of 320 pairs of words, were given to 41 members of a class in introductory psychology. The time allowed was five minutes for each list of 80 pairs, but in each case, the subjects were stopped after two and a half minutes and asked to indicate the point reached. They were then allowed to continue for two and a half minutes longer. It was thought that a total time allowance of five minutes for each list would be sufficient to enable the majority of subjects to complete the whole list, and this was found to be true.

Two months after the first test, 33 members of the same group were retested with List I of the original series of four lists, under

the same conditions, using the same directions and with the same time allowance. Three of the records were discarded because of the large proportion of omissions. This group of 30 will be referred to as Group I; the first of the four lists which was later repeated will be called Test I, and the repetition of it, Test II.

A comparison of the results of Tests I and II is given in table 25. The average number of correct responses as well as the average score "right minus wrong," is higher in Test II than in Test I. It is thought that this increase is probably due more to habituation to the type of response demanded by the test than to practice in the actual pairs of words although it may be a combined effect of the two factors. It will be remembered that Test I was followed immediately by 240 additional pairs of words, and Test II was given after an interval of two months, so that it does not seem

TABLE 25

Comparison of the average and standard deviation of the number of correct responses and the score in Test I and Test II

	2½ MINUTES				5 MINUTES			
	Number right		Score		Number right		Score	
	Aver-age	S.D.	Aver-age	S.D.	Aver-age	S.D.	Aver-age	S.D.
Test I	29.5	13.3	19.2	16.1	56.4	14.7	41.4	19.6
Test II	38.0	14.2	26.6	17.8	63.5	9.8	48.1	18.2

likely that many of the first 80 pairs of words responded to were remembered. It would probably have been a better procedure to use the second or third list for repetition, instead of the first list, so that the practice effect, if there was one, could have been somewhat reduced. As the four lists were not known to be equivalent, nothing could be concluded about practice effects in this instance.

The coefficient of correlation between the scores in that part of the test completed in two and a half minutes in Tests I and II, is 0.859; between the number of correct responses in the same part of the tests, 0.851. It seems, therefore, that one measure is just about as reliable as the other. This indicates nothing, however, concerning the validity of either measure as a true measure of an individual's performance.

A comparison of the responses to individual pairs of stimulus words is of more interest. Only the stimulus words which were responded to in both tests, during the total time of five minutes are considered. The means and standard deviations for the number of correct responses and the score in both tests are given in table 26. Both the average number of correct responses and the average score are slightly higher in Test II than in Test I. If there had been no practice effect of any sort, it is assumed that the averages would have been the same for both tests.

The coefficient of correlation between the number right in the two tests is 0.966 (as compared with 0.851 when the number right in two and a half minutes was considered). This is only slightly higher than the correlation between scores in the two tests, which is 0.930 (as compared with 0.859 for the scores in the two and

TABLE 26

Comparison of the average and standard deviation of the number of correct responses and the score, for the words responded to in both Test I and Test II by 30 members of Group I

	NUMBER CORRECT RESPONSES		SCORE (RIGHT-WRONG)	
	Average	S.D.	Average	S.D.
Test I	56.40	14.73	41.10	20.08
Test II	57.77	14.28	43.93	19.64

one-half minutes). There is a correlation of 0.954 between the number right and the score in Test I, and of 0.947 in Test II. These coefficients are so nearly equal that one measure may be considered as reliable as another, but here again, the information derived thus far does not contribute anything to the question of the validity of either measure.

As we have stated before, in the type of test in which there are only two possible responses, that is, in a "true-false" test, it is assumed according to the law of chance, that of the total number of responses not known, one-half will be answered incorrectly and one-half correctly. In a repetition of the same test, assuming that a subject has not learned any of the responses not known at the time of the first test, the number of errors should be the same for both tests, but the actual items wrong will

TABLE 27

An analysis of the errors for individual cases in Group I, in Tests I and II

CASE	TOTAL ERRORS	ERRORS IN TEST I	ERRORS IN TEST II	ERRORS IN TEST I NOT IN TEST II	ERRORS IN TEST II NOT IN TEST I	ERRORS IN BOTH TESTS	PER CENT TOTAL ERRORS IN BOTH TESTS
1	33	18	15	10	7	8	24.1
2	49	22	27	6	11	16	32.6
3	24	14	10	11	7	3	12.5
4	22	12	10	5	3	7	31.8
5	40	23	17	10	4	13	32.5
6	37	22	15	11	4	11	28.9
7	17	11	6	7	2	4	23.5
8	12	7	5	4	2	3	25.5
9	6	3	3	1	1	2	33.3
10	32	16	16	6	6	10	31.2
11	66	32	34	11	13	21	31.8
12	18	14	4	12	2	2	11.1
13	26	12	14	3	5	9	34.5
14	13	5	8	2	5	3	23.0
15	26	14	12	6	4	8	30.7
16	55	26	29	7	10	19	34.5
17	26	14	12	7	5	7	26.9
18	30	15	15	9	9	6	20.0
19	37	19	18	12	11	7	18.9
20	44	22	22	11	10	12	27.2
21	26	13	13	5	5	8	30.7
22	27	14	13	6	5	8	29.6
23	33	19	14	9	4	10	30.3
24	7	4	3	2	1	2	28.5
25	38	18	20	8	10	10	26.3
26	16	10	6	7	3	3	18.7
27	15	7	8	2	3	5	33.3
28	18	8	10	1	3	7	38.8
29	16	6	10	2	6	4	25.0
30	57	31	26	14	9	17	29.8
Total.....	867	452	415	207	170	245	

not be identical in both tests. If x equals the total number of errors in both tests together, $\frac{x}{2}$ equals the number of errors in each of the two tests. Of $\frac{x}{2}$ errors in Test I, $\frac{1}{2}$ will be identical

in Test II, that is, $\frac{1}{4}$, or 25 per cent of the total number of errors in both tests will be common to both tests. In table 27 are given the results for each of the thirty individuals in this group. The percentages of total errors common to both tests show a wide individual variation, ranging from 11.1 per cent to 38.8 per cent.

A comparison of the actual and theoretical cases based on the total number of errors for all 30 cases as found in this group is given in table 28. Of a total of 867 errors in the two tests, 452 appear in Test I and 415 in Test II. Of the 452 errors in Test I, 207 are found only in Test I and the remaining 245 are common to both tests; of the 415 errors in Test II, 170 are found only in Test II, and 245 are in Test I also. That is, out of 867 errors, 245 are

TABLE 28

Actual distribution of errors in Test I and Test II for Group I, and the theoretical or chance distribution based upon the total number of errors in Tests I and II

	ACTUAL CASE	THEORETICAL CASE
Total errors in both tests	867	867
Errors in test I	452	433.5
Errors in test II	415	433.5
Errors in test I but not in test II	207	216.75
Errors in test II but not in test I	170	216.75
Errors common to both tests	245	216.75
Per cent of total errors common to both tests	28.25	25.00

common to both tests. Or, in other words, 245 wrong responses given in Test I are matched word for word in Test II.

Thus, instead of 25 per cent of the total number of errors common to both tests, the actual per cent found is 28.35 per cent. This indicates that there is some factor other than chance, operating at least to a small degree in determining the responses to the total number of items not known. The individual variation has been seen to be considerable. It may be that some of the subjects have very definite misconceptions about the relations of some of the pairs of words so that, for example, they mark two synonyms as antonyms quite deliberately and repeat the error in both tests. There may be certain stimulus words which lend themselves to this sort of misconception more readily than others. Some

light will be thrown upon this question in the study of the responses to Group II.

There is a suggestion from the findings thus far, that on the average, the "R-W" method of scoring gives a score which represents a somewhat smaller number of items than is actually known by the subjects.

IV. COMPARISON OF RESPONSES IN TWO TYPES OF SYNONYM-ANTONYM TEST

As a part of a combination test administered to 223 entering students at the Johns Hopkins University in October, 1923, the "synonym-antonym" test, equivalent to Anderson's Form I list, was given, with the substitution of the following directions in place of the directions as given by Anderson:

> The two words in each of the following pairs of words are either similar or opposite in meaning. If the two words of a pair are similar in meaning, put an "S" on the line between them. If they are opposite in meaning, put an "O" on the line between them. The samples are correctly marked.
>
> Sample: white O black
> big S large

The time allowed was two and a half minutes. This test was given first in the group of tests and was followed by 32.5 minutes of work on other tests.

Two months later, 150 members of the original group, in an English Composition class were retested with a synonym-antonym test of another form.

Twenty-seven pairs of antonyms and 27 pairs of synonyms from the list as given in October, were combined with 27 pairs of words neither similar nor opposite in meaning, so that 54 pairs of words from the first test were distributed in a chance order, through the total of 81 pairs of words in Test II. A similar list was made up including 54 pairs of words from Anderson's second list (fig. 7).

The stimulus words were arranged in a column to the left of the page. To the right of each pair of stimulus words, were written

FIG. 7

If the two words of a pair are similar in meaning, underline the word Similar in the list of words to the right of the pair. If the two words of a pair are opposite in meaning, underline the word Opposite. If they are neither similar nor opposite in meaning, underline the word Neither. The samples are correctly marked.

Samples: 1. black	white	Similar, Opposite, Neither
2. tired	weary	Similar, Opposite, Neither
3. empty	soft	Similar, Opposite, Neither

Test 1

1. masculine	feminine	1. Similar, Opposite, Neither
2. acquire	lose	2. Similar, Opposite, Neither
3. discrepancy	invective	3. Similar, Opposite, Neither
4. mastery	ascendancy	4. Similar, Opposite, Neither
5. vanity	conceit	5. Similar, Opposite, Neither
6. benign	accidental	6. Similar, Opposite, Neither
7. default	defalcation	7. Similar, Opposite, Neither
8. anxiety	nonchalance	8. Similar, Opposite, Neither
9. node	embezzlement	9. Similar, Opposite, Neither
10. decadence	decline	10. Similar, Opposite, Neither
11. merriment	gloom	11. Similar, Opposite, Neither
12 extant	ambiguous	12. Similar, Opposite, Neither
13. knave	villain	13. Similar, Opposite, Neither
14. congruity	diatribe	14. Similar, Opposite, Neither
15. greediness	cupidity	15. Similar, Opposite, Neither
16. accountable	irresponsible	16. Similar, Opposite, Neither
17. genial	adventitious	17. Similar, Opposite, Neither
18. immune	susceptible	18. Similar, Opposite, Neither
19. adapt	conform	19. Similar, Opposite, Neither
20. extinct	equivocal	20. Similar, Opposite, Neither
21. momentous	immaterial	21. Similar, Opposite, Neither
22. specific	general	22. Similar, Opposite, Neither
23. pompous	ostentatious	23. Similar, Opposite, Neither
24. knot	peculation	24. Similar, Opposite, Neither
25. disgust	aversion	25. Similar, Opposite, Neither
26. waste	conserve	26. Similar, Opposite, Neither
27. erudite	accessory	27. Similar, Opposite, Neither
28. any	none	28. Similar, Opposite, Neither
29. desultory	rambling	29. Similar, Opposite, Neither
30. defile	purify	30. Similar, Opposite, Neither
31. ghost	substance	31. Similar, Opposite, Neither
32. avert	prevent	32. Similar, Opposite, Neither
33. scholarly	foreign	33. Similar, Opposite, Neither
34. putrid	droll	34. Similar, Opposite, Neither
35. vestige	trace	35. Similar, Opposite, Neither
36. fetid	esoteric	36. Similar, Opposite, Neither
37. acme	climax	37. Similar, Opposite, Neither
38. motile	germane	38. Similar, Opposite, Neither
39. accumulate	dissipate	39. Similar, Opposite, Neither
40. obedient	refractory	40. Similar, Opposite, Neither

FIG. 7—*Continued*

41. moderate	violent	41. Similar, Opposite, Neither
42. apprehensive	fearful	42. Similar, Opposite, Neither
43. traditional	legendary	43. Similar, Opposite, Neither
44. exoteric	sessile	44. Similar, Opposite, Neither
45. myopia	nescience	45. Similar, Opposite, Neither
46. hasty	circumspect	46. Similar, Opposite, Neither
47. hoax	deception	47. Similar, Opposite, Neither
48. synchronous	simultaneous	48. Similar, Opposite, Neither
49. amiable	surly	49. Similar, Opposite, Neither
50. abjure	append	50. Similar, Opposite, Neither
51. blot out	delete	51. Similar, Opposite, Neither
52. abettor	eulogy	52. Similar, Opposite, Neither
53. suave	brusque	53. Similar, Opposite, Neither
54. encomium	conflict	54. Similar, Opposite, Neither
55. superfluous	essential	55. Similar, Opposite, Neither
56. irregular	eccentric	56. Similar, Opposite, Neither
57. concur	hint	57. Similar, Opposite, Neither
58. vilify	praise	58. Similar, Opposite, Neither
59. apathy	indifference	59. Similar, Opposite, Neither
60. previous	succeeding	60. Similar, Opposite, Neither
61. prefix	renounce	61. Similar, Opposite, Neither
62. wax	prosy	62. Similar, Opposite, Neither
63. pertinacious	obstinate	63. Similar, Opposite, Neither
64. proclivity	inclination	64. Similar, Opposite, Neither
65. formulate	palliate	65. Similar, Opposite, Neither
66. recant	disavow	66. Similar, Opposite, Neither
67. contradict	corroborate	67. Similar, Opposite, Neither
68. mitigate	express	68. Similar, Opposite, Neither
69. dignify	degrade	69. Similar, Opposite, Neither
70. ligature	band	70. Similar, Opposite, Neither
71. gaunt	buxom	71. Similar, Opposite, Neither
72. hyperopia	knowledge	72. Similar, Opposite, Neither
73. hypothesis	surmise	73. Similar, Opposite, Neither
74. martial	civil	74. Similar, Opposite, Neither
75. amenable	tractable	75. Similar, Opposite, Neither
76. vesper	matin	76. Similar, Opposite, Neither
77. imply	wane	77. Similar, Opposite, Neither
78. even	protuberant	78. Similar, Opposite, Neither
79. contingent	dependent	79. Similar, Opposite, Neither
80. sarcastic	hopeless	80. Similar, Opposite, Neither
81. scanty	docile	81. Similar, Opposite, Neither
1. often	seldom	1. Similar, Opposite, Neither
2. con	pro	2. Similar, Opposite, Neither
3. demonstrate	execrate	3. Similar, Opposite, Neither
4. largess	donation	4. Similar, Opposite, Neither
5. degradation	humiliation	5. Similar, Opposite, Neither
6. decrease	revile	6. Similar, Opposite, Neither
7. indict	arraign	7. Similar, Opposite, Neither
8. depressed	elated	8. Similar, Opposite, Neither
9. deviation	suavity	9. Similar, Opposite, Neither
10. radiance	effulgence	10. Similar, Opposite, Neither

FIG. 7—*Continued*

11. vagrant	dweller	11. Similar, Opposite, Neither
12. fallacy	asperity	12. Similar, Opposite, Neither
13. orifice	aperture	13. Similar, Opposite, Neither
14. pertinent	penurious	14. Similar, Opposite, Neither
15. linger	loiter	15. Similar, Opposite, Neither
16. transient	permanent	16. Similar, Opposite, Neither
17. relevant	avaricious	17. Similar, Opposite, Neither
18. irksome	refreshing	18. Similar, Opposite, Neither
19. recoup	recover	19. Similar, Opposite, Neither
20. representative	meticulous	20. Similar, Opposite, Neither
21. actual	fictitious	21. Similar, Opposite, Neither
22. facility	difficulty	22. Similar, Opposite, Neither
23. confer	grant	23. Similar, Opposite, Neither
24. proxy	perfunctory	24. Similar, Opposite, Neither
25. plenary	complete	25. Similar, Opposite, Neither
26. integrity	dishonesty	26. Similar, Opposite, Neither
27. profligate	choleric	27. Similar, Opposite, Neither
28. besmirch	cleanse	28. Similar, Opposite, Neither
29. obdurate	stubborn	29. Similar, Opposite, Neither
30. significant	paltry	30. Similar, Opposite, Neither
31. discretion	folly	31. Similar, Opposite, Neither
32. lucrative	profitable	32. Similar, Opposite, Neither
33. sable	phlegmatic	33. Similar, Opposite, Neither
34. brevity	apathy	34. Similar, Opposite, Neither
35. tedious	dry	35. Similar, Opposite, Neither
36. purity	feeling	36. Similar, Opposite, Neither
37. torpor	stupor	37. Similar, Opposite, Neither
38. celibate	evanescent	38. Similar, Opposite, Neither
39. imprisoned	free	39. Similar, Opposite, Neither
40. carnivorous	herbivorous	40. Similar, Opposite, Neither
41. aggrandize	belittle	41. Similar, Opposite, Neither
42. influence	incentive	42. Similar, Opposite, Neither
43. facilitate	further	43. Similar, Opposite, Neither
44. stubborn	ephemeral	44. Similar, Opposite, Neither
45. loquacious	lustrous	45. Similar, Opposite, Neither
46. insipidity	zest	46. Similar, Opposite, Neither
47. scarcity	dearth	47. Similar, Opposite, Neither
48. null	void	48. Similar, Opposite, Neither
49. sterile	fertile	49. Similar, Opposite, Neither
50. vindictive	metaphysical	50. Similar, Opposite, Neither
51. avarice	cupidity	51. Similar, Opposite, Neither
52. temporary	intrepid	52. Similar, Opposite, Neither
53. terrestrial	celestial	53. Similar, Opposite, Neither
54. enduring	gallant	54. Similar, Opposite, Neither
55. convoke	dismiss	55. Similar, Opposite, Neither
56. mysterious	cryptic	56. Similar, Opposite, Neither
57. supercilious	continual	57. Similar, Opposite, Neither
58. straight	tortuous	58. Similar, Opposite, Neither
59. latent	hidden	59. Similar, Opposite, Neither
60. effeminate	virile	60. Similar, Opposite, Neither

FIG. 7—*Concluded*

61. disdainful	interminable	61. Similar, Opposite, Neither
62. forthwith	recondite	62. Similar, Opposite, Neither
63. infrequently	seldom	63. Similar, Opposite, Neither
64. assiduous	diligent	64. Similar, Opposite, Neither
65. eventually	abstruse	65. Similar, Opposite, Neither
66. incumbent	obligatory	66. Similar, Opposite, Neither
67. adversary	colleague	67. Similar, Opposite, Neither
68. agglomerate	finite	68. Similar, Opposite, Neither
69. manacle	extricate	69. Similar, Opposite, Neither
70. fractious	irritable	70. Similar, Opposite, Neither
71. comprehensive	restricted	71. Similar, Opposite, Neither
72. scatter	maxim	72. Similar, Opposite, Neither
73. desuetude	disuse	73. Similar, Opposite, Neither
74. altruistic	egotistic	74. Similar, Opposite, Neither
75. innuendo	insinuation	75. Similar, Opposite, Neither
76. impecunious	opulent	76. Similar, Opposite, Neither
77. peroration	urbanity	77. Similar, Opposite, Neither
78. lead	follow	78. Similar, Opposite, Neither
79. intrigue	cabal	79. Similar, Opposite, Neither
80. civility	aphorism	80. Similar, Opposite, Neither
81. limited	preamble	81. Similar, Opposite, Neither

the words, "similar, opposite, neither." The 162 pairs of words were arranged on four sheets. The directions were as follows:

If the two words of a pair are similar in meaning, underline the word Similar in the list of words to the right of the pair. If the two words of a pair are opposite in meaning, underline the word Opposite. If they are neither similar nor opposite in meaning, underline the word Neither. The samples are correctly marked. For example, "Black" and "White" are opposite in meaning, therefore the word "Opposite" is underlined. "Tired" and "Weary" are similar in meaning, therefore the word "Similar" is underlined. "Empty" and "Soft" are neither similar nor opposite in meaning, therefore the word "Neither" is underlined.

As one of the purposes of the retest was to compare the responses to individual pairs of words in the two forms of the test, it was desirable that sufficient time be allowed for the completion of the first 81 pairs of words, which included all pairs of words common to both lists. After some preliminary trial, it was decided to allow 6 minutes. This time allowance proved to be too short for all of the subjects to complete the first 81 pairs of words.

The 150 individuals for whom there are records in both tests will be designated Group II. The synonym-antonym test as given in October will be referred to as Test I, and the synonym-antonym test with the "neither" words included, as Test II.

Comparison of responses in the two tests

Since the time allowed for Test I was two and a half minutes, and that allowed for Test II, six minutes, it is not possible to draw a direct comparison between the gross results in each test. In table 29, the average number of items attempted, the average number omitted, wrong and right are given for both tests. It is assumed in determining the number of items attempted that the subject has worked through the last item checked. The number of items attempted in Test II naturally exceeds that for Test I. The number of omissions is particularly large for Test II and it is probably true that in some instances, the subject did not attempt a judgment for each pair of words omitted, but ran hurriedly through the list checking only the obvious pairs. The increase in number of items attempted in Test II is not proportionate to

TABLE 29

Comparison of results in Test I and Test II for the same 150 cases

	TIME ALLOWED	AVERAGE NUMBER OF ITEMS "ATTEMPTED"	AVERAGE NUMBER OMITTED	AVERAGE NUMBER WRONG	AVERAGE NUMBER RIGHT
	minutes				
Test I	$2\frac{1}{2}$	44.2	2.0	5.5	36.6
Test II	6	74.8	6.1	14.3	54.7

the increase in time allowed which brings out one fact, *i.e.*, that more time is required to make a judgment when there are three categories to choose from than when there are only two.

Analysis of responses to stimulus words in the synonym-antonym test including the "neither" category

The types of response in the synonym-antonym test without the "neither" group have already been discussed. It was found that the percentage of accuracy was, on the average, higher for antonyms than for synonyms.

The responses for the 150 individuals in Group II, to the first 81 pairs of words in Test II will next be studied. An analysis of the responses is given in table 30. Had each individual in the group completed all 81 responses, the possible number of re-

sponses would have been equal for the three types of pairs of words, synonyms, antonyms and "neithers." Of the 11,229 actual responses there are 38 fewer "neither" responses than antonyms, and 11 fewer antonyms than synonyms. These differences are too slight to affect the results. The antonym group shows the smallest number of errors and omissions, and, consequently, the largest number of correct responses. The antonyms also have a higher percentage of accuracy than the synonyms in Test I. It is interesting, however, that the synonyms show a higher percentage of accuracy than the "neither" group. The number of omissions is more than twice as great for the

TABLE 30

Analysis of responses to the three types of stimulus words in Test II

	SYNONYMS	ANTONYMS	"NEITHER"	TOTAL
Possible number of responses including omissions	3763	3752	3714	11229
Number omitted	204	121	589	914
Number of responses excluding omissions	3559	3631	3125	10315
Number wrong	785	596	724	2105
Number right	2774	3035	2401	8210
Number of times each response was made	3379	3648	3288	10315
Number of times used, when wrong	605	613	887	2105
Average per cent of accuracy	72.7	79.9	65.8	72.8

"neither" group than for either synonyms or antonyms. If the percentage of accuracy is based on the actual number of responses made (leaving out of account the omissions), the relative position of the three groups does not change, that is, the antonyms have the highest percentage and the "neither" group the lowest percentage of accuracy. There is a possibility, although it is purely an inference, that some of the subjects thought that the "neither" group was introduced as a catch and avoided making "neither" judgments. It was found that the actual number of "neither" judgments was smaller than the number of either "similar" or "opposite" judgments.

Since the relative difficulty of the individual words in the pairs of stimulus words has not been determined, it is impossible

to draw inferences from the data thus far presented as to the relative difficulty of the processes involved in judging the relationships between the words. Data will be presented as to the relative difficulty of pairs of words, as determined by the percentage of accuracy, but they give only indirect evidence as to the words themselves.

An analysis of the errors made in each of the three types of pairs of words is presented in tables 31, 32 and 33. In table 31, the 27 pairs of synonyms are listed, with the per cent of omissions, errors and correct responses for each pair. The errors are further analyzed into "neither" and "opposite" responses. The pairs of words are arranged in accordance with this analysis. That is, those synonyms for which the errors are predominately of the "opposite" type are listed first, followed by those which show errors, predominately of the "neither" type. If the percentage of accuracy of the synonyms showing a greater proportion of errors of the "opposite" type are averaged separately from those of the synonyms showing errors predominately of the "neither" type, it becomes evident that the "easier" synonyms, as judged by the percentage of accuracy, are more often judged "opposite" than "neither" in meaning by those who make errors. The more difficult synonyms are more often judged to be "neither" than "opposite." There is a tendency, therefore, for individuals to recognize a definite relationship between words, or rather, to perceive that the words are related either as "opposite" or "similar," rather than "neither" in meaning. Either confusion or carelessness leads them to make the wrong response. There is an equally distinct tendency for some synonyms to be definitely judged as neither similar nor opposite in meaning. If it were true that pairs of words that are not familiar to the subject are marked simply by chance, an equal number of "opposite" and "neither" responses would be expected. If the percentage of "neither" and "opposite" responses are averaged for all 27 synonyms the average of "neither" responses which is 59.8 exceeds the average of "opposite" responses which is 40.1, by a difference of 9.79 points.

A similar analysis of the 27 antonyms is given in table 32. Those antonyms are listed first in which the errors are more

TABLE 31

Distribution of responses to synonyms

SYNONYMS	NUMBER OF RESPONSES	OMITTED	WRONG	RIGHT	WRONG JUDGED "OPPOSITE"	WRONG JUDGED "NEITHER"	AVERAGE PER CENT
		per cent	*per cent*	*per cent*	*per cent*	*per cent*	
Traditional-legendary	147	1	5	94	88	12	
Recant-disavow	128	13	21	66	70	30	
Knave-villain	150	1	11	88	69	31	
Vanity-conceit	150	1	13	86	68	32	Omitted, 4.8
Apathy-indifference	143	8	33	59	64	36	Wrong, 18.5
Irregular-eccentric	145	2	30	68	60	40	Right, 76.4
Hypothesis-surmise	106	4	24	72	60	40	
Proclivity-inclination	132	13	23	64	58	42	
Avert-prevent	150	1	7	92	55	45	
Hoax-deception	147	2	8	90	50	50	
Synchronous-simultaneous	146	9	29	62	48	52	
Pertinacious-obstinate	135	6	19	75	48	52	
Disgust-aversion	150	3	27	71	40	60	
Desultory-rambling	150	10	21	69	34	66	
Amenable-tractable	94	17	19	64	33	67	
Contingent-dependent	83	8	47	45	28	72	
Greediness-cupidity	150	5	34	61	27	73	
Apprehension-fearful	147	2	18	80	26	74	Omitted, 6.4
Mastery-ascendancy	150	6	29	65	25	75	Wrong, 23.8
Adapt-conform	150	1	12	87	22	78	Right, 69.7
Decadence-decline	150	2	13	85	20	80	
Pompous-ostentatious	150	3	17	80	20	80	
Default-defalcation	150	10	25	65	18	82	
Blot out-delete	146	5	8	87	18	82	
Vestige-trace	150	3	21	76	13	87	
Acme-climax	150	8	25	67	11	89	
Ligature-band	114	12	41	47	11	89	
Average		5.9	21.4	72.7	40.14	59.83	

TABLE 32

Distribution of responses to antonyms

ANTONYMS	NUMBER OF RESPONSES	OMITTED	WRONG	RIGHT	WRONG JUDGED "SIMILAR"	WRONG JUDGED "NEITHER"	AVERAGE PER CENT
		per cent	*per cent*	*per cent*	*per cent*	*per cent*	
Waste-conserve	150	0	2	98	100	0	
Specific-general	150	0	9	91	85	15	
Previous-succeeding	142	0	20	80	82	18	
Immune-susceptible	150	2	13	85	75	25	
Vesper-matin	92	15	34	51	71	29	Omitted, 4.1
Contradict-corroborate	124	5	10	85	62	38	Wrong, 14.7
Martial-civil	104	3	14	83	60	40	Right, 81.0
Gaunt-buxom	111	9	20	71	59	41	
Any-none	150	2	16	82	58	42	
Amiable-surly	146	2	8	90	58	42	
Suave-brusque	145	8	16	76	57	43	
Accountable-irresponsible	150	1	8	91	50	50	
Vilify-praise	144	10	17	73	50	50	
Anxiety-nonchalance	150	5	13	82	35	65	
Superfluous-essential	145	3	14	83	30	70	
Hasty-circumspect	147	6	46	48	16	84	
Even-protuberant	85	20	25	55	15	85	
Defile-purify	150	0	9	91	14	86	
Ghost-substance	150	1	33	66	12	88	
Dignify-degrade	120	2	13	85	12	88	Omitted, 3.3
Obedient-refractory	149	4	25	71	11	89	Wrong, 19.4
Momentous-immaterial	150	1	35	64	10	90	Right, 77.0
Accumulate-dissipate	150	2	21	77	3	97	
Acquire-lose	150	0	2	98	0	100	
Merriment-gloom	150	0	1	99	0	100	
Moderate-violent	148	0	16	84	0	100	
Masculine-feminine	150	0	0	100	0	0	
Average		3.7	16.2	79.9	39.3	60.5	

TABLE 33

Distribution of responses to "neither" words

"NEITHER"	NUMBER OF RESPONSES	OMITTED	WRONG	RIGHT	WRONG JUDGED 'SIMILAR'	WRONG JUDGED 'OPPOSITE'	AVERAGE PER CENT
		per cent	*per cent*	*per cent*	*per cent*	*per cent*	
extant-ambiguous	150	8	31	61	85	15	
mitigate-express	124	9	38	53	85	15	
scanty-docile	78	2	8	90	83	17	
discrepancy-invective	150	18	15	67	78	22	
knot-peculation	150	16	19	65	75	25	
concur-hint	144	7	31	62	73	27	
hyperopia-knowledge	109	22	18	60	70	30	Omitted, 15.2
putrid-droll	150	8	33	59	68	32	Wrong, 23.6
erudite-accessory	150	16	13	71	65	35	Right, 61.0
genial-adventitious	150	14	31	55	64	36	
node-embezzlement	150	21	10	69	60	40	
fetid-esoteric	150	25	25	50	59	41	
formulate-palliate	130	26	29	45	58	42	
abjure-append	146	22	30	48	55	45	
wax-prosy	136	14	7	79	50	50	
imply-wane	91	12	13	75	50	50	
benign-accidental	150	8	16	76	42	58	
encomium-conflict	145	20	20	60	38	62	
myopia-nescience	147	35	26	39	37	63	
extinct-equivocal	150	10	13	77	35	65	
sarcastic-hopeless	82	1	4	95	33	67	Omitted, 15.2
motile-germane	150	29	19	52	28	72	Wrong, 14.8
abettor-eulogy	146	11	12	77	28	72	Right, 69.9
congruity-diatribe	150	23	17	60	24	76	
exoteric-sessile	147	28	18	54	23	77	
prefix-renounce	139	3	9	88	15	85	
scholarly-foreign	150	0	9	91	0	100	
Average		15.1	19.0	65.8	51.1	48.8	

frequently of the "similar" type, followed by those in which the errors are predominately of the "neither" type. The average percentage of accuracy for the eleven antonyms which show a greater proportion of errors of the "similar" type is 81.0 as compared with 77.0 for the antonyms showing a larger proportion of errors of the "neither" type. That is, here again, as was true of the synonyms, there is a group of antonyms which are, apparently, familiar enough to suggest a definite relationship to the subject, but the relationship is often confused. To illustrate with a concrete case, 20 per cent of the responses to as familiar antonyms as "preceding and succeeding" are errors. Of this 20 per cent, 82 per cent are of the "opposite" type and only 18 per cent of the "neither" type. It would seem that a subject in perceiving this pair of words, eliminates immediately the "neither" category, but becomes confused between the "similar" and "opposite" categories.

There are as definite tendencies to judge antonyms as "neither," similar nor opposite in meaning. To illustrate, 35 per cent of the total number of responses to the antonyms "momentous-immaterial" are errors. Of this 35 per cent, 90 per cent are "neither" judgments and only 10 per cent "similar" judgments.

The average of the per cent of errors of the "similar" type for all 27 pairs of antonyms is 39.3, which is exceeded by the average of the per cent of errors of the "neither" type, 60.5, by 21.2 points. That is, on an average, there is less tendency for errors in antonyms to be of the "similar" type than of the "neither" type.

An analysis of the 27 pairs of "neither" words is presented in table 33. The pairs of words are listed first which yield a larger proportion of errors due to their being judged as "similar" in meaning, followed by those showing errors predominately of the "opposite" type. To illustrate, 31 per cent of the responses to the words "extant-ambiguous" are incorrect. Of this 31 per cent, 85 per cent of the responses are "similar" and 15 per cent "opposite." Eighteen per cent of the responses to "exoteric-sessile" are incorrect. Of these incorrect responses, 77 per cent are "opposite" and 23 per cent "similar." Of the 9 per cent of incorrect responses to "scholarly-foreign," 100 per cent are "opposite." The average percentage of accuracy is higher for the

"neither" words judged most frequently to be "opposite" than for those judged to be "similar" more often. The total number of errors for all 27 pairs of "neither" words is divided almost evenly between "similar" and "opposite" responses.

It seems clear thus far that relationships, other than the correct ones, are definitely suggested by some of the pairs of words. We are not dealing with a situation in which the exact relationship between pairs of words is either known or not known. It is possible that some of the pairs of words suggest relationships to some of the subjects which seem entirely logical to them. It would be interesting to discover, for example, why 39 out of 150 individuals judged "extant-ambiguous" to be similar in meaning, whereas only 7 out of the same 150 judged them to be opposite in meaning.

Comparison of the responses to the stimulus words responded to in both forms of synonym-antonym test; the effect of the "neither" category

In order to determine more definitely the effect of including a "neither" group with synonyms and antonyms, the responses given to the 54 pairs of words in both tests were compared. Test I was composed entirely of synonyms and antonyms; Test II included a "neither" group, together with the synonyms and antonyms. For a given individual, only those pairs of words are considered in this comparison which were responded to in both tests, excluding all omissions. Thus there are 6019 responses, correct and incorrect in each test, which will be compared. The number of responses to some of the pairs of words is considerably less than 150.

The coefficient of correlation between the number of correct responses to the stimulus words responded to in both tests is 0.926.

An attempt has been made to ascertain what differences in response in the two tests chance alone would account for, before drawing any conclusions as to the actual differences found to exist. As has been stated in another connection, the assumption is made that of the total number of responses not known to the subject, but guessed at, given proportions will be checked

right and wrong by chance. If there are only two possible responses, it is assumed that according to chance, 50 per cent of the total group not actually known by the subject will be checked correctly and 50 per cent incorrectly. If there are three possible responses, only one of which is correct, 67 per cent of the total number not known will be checked incorrectly and 33 per cent correctly. Test I is of the first type, and Test II of the second type. It is necessary, then, to determine the percentage of errors that may be expected in Test II on the basis of the errors in Test I. It seems logical to assume that the ratio of the per cent of errors in Test I to the per cent of errors in Test II should be as 50 to 67.

The actual number and per cent of wrong and right responses in both Tests I and II are given in table 34. Of the total number

TABLE 34

Comparison of responses to the stimulus words responded to in both Test I and Test II

	TOTAL NUMBER RESPONDED TO	NUMBER WRONG	NUMBER RIGHT	PER CENT WRONG	PER CENT RIGHT
Test I	6,019	717	5,302	11.91	88.08
Test II	6,019	992	5,027	16.47	83.52

of responses in each test, 11.91 per cent in Test I and 16.47 per cent in Test II are wrong. It would have been predicted from the per cent of errors in Test I that 15.95 per cent of responses in Test II would be incorrect, since 50:67::11.9:15.95. Thus, without considering the responses to any of the individual stimulus words, and averaging together the responses of all of the subjects, it would appear that there is some justification for the assumption, already stated, concerning the checking of items not actually known by the subject. But an analysis of the actual pairs of words reveals a different sort of situation.

It seems fair to make the same assumption with regard to the individual pairs of words that was made for all of the responses for a given subject. That is, it can be assumed that of the total number of subjects not knowing the correct response to a given pair of words, one-half will give the correct response, and one-

half the incorrect response, purely by chance, if there are only two possible responses; but if there are three possible responses two-thirds will be incorrect and one-third correct. The percentage of incorrect responses for a given pair of words in Test I should be in the ratio of 50 to 67 to the percentage of incorrect responses for the same pair of words in Test II. The per cent of errors for the synonyms and antonyms in Test I and Test II, and the differences between the two tests, together with the number of cases on which the per cents are based are shown in tables 35 and 36. The synonyms are listed in table 35 and the antonyms in table 36.

An examination of the synonyms, first, reveals the fact that for individual pairs of words, it cannot be predicted what the response will be in Test II on the basis of the response in Test I. Some of the synonyms show a higher percentage of accuracy and some a lower, when the "neither" group is included, whereas if chance alone were operating in the checking of responses not known, the per cent of accuracy should be lower for all pairs of words in Test II. The synonyms showing a higher percentage of accuracy in Test II (indicated by minus differences in table 35) are, with two exceptions, to be found in table 31, in the group of synonyms which, when judged incorrectly, are more frequently judged as opposite in meaning, instead of as "neither;" that is, as was suggested before, they are probably the more familiar synonyms, which are judged incorrectly through carelessness, or temporary confusion. The synonyms which are much less accurately judged in Test II, when the "neither" group is included (indicated by large positive differences in table 35) are all among those judged to be "neither" more frequently than "opposite." The effect of introducing the "neiteer" group of words, is, then, to make the judgment of the relationship existing between some pairs of words easier, but more difficult for other pairs of words.

An examination of the antonyms in table 36 leads to the same general conclusions. Here, again, the antonyms which are judged more accurately in Test II than in Test I, are among the antonyms (in table 32) which when incorrectly judged, are more frequently judged to be "similar" in meaning, and these are the "familiar" antonyms. Those judged much less accurately in

Test II, are the ones which were judged more frequently to be "neither" in meaning.

These results for both synonyms and antonyms indicate the differences in difficulty presented by different stimuli. For some

TABLE 35

Per cent of errors in Test I and Test II

	SYNONYMS	NUMBER OF CASES	PER CENT WRONG IN TEST I	PER CENT WRONG IN TEST II	DIFFERENCE BETWEEN TEST I AND TEST II
4	mastery-ascendancy	133	23	31	8
5	vanity-conceit	146	17	13	−4
7	default-defalcation	137	12	25	13
10	decadence-decline	144	10	13	3
13	knave-villain	148	14	12	−2
15	greediness-cupidity	140	28	35	7
19	adapt-conform	141	6	11	5
23	pompous-ostentatious	138	15	18	3
25	disgust-aversion	142	23	26	3
29	desultory-rambling	127	20	23	3
32	avert-prevent	147	11	7	−4
35	vestige-trace	139	16	20	4
37	acme-climax	130	15	25	10
42	apprehensive-fearful	138	19	16	−3
43	traditional-legendary	136	6	5	−1
47	hoax-deception	128	5	8	3
48	synchronous-simultaneous	20	20	25	5
51	blot out-delete	117	7	7	0
56	irregular-eccentric	120	14	27	13
59	apathy-indifference	96	22	35	13
63	pertinacious-obstinate	97	20	20	0
64	proclivity-inclination	67	19	16	−3
66	recant-disavow	60	23	22	−1
70	ligature-band	52	12	21	9
73	hypothesis-surmise	48	17	15	−2
75	amenable-tractable	37	27	14	−13
79	contingent-dependent	25	32	48	16
	Average	109.3	16.7	19.9	3.1

pairs of words, a subject probably guesses between only two of the three possible categories, whereas for other pairs of words, he guesses between all three categories. If this is true then, obviously, it is not possible to take account of the chance factor

since it is altogether unmeasurable and we are not justified in subtracting a certain proportion of the errors from the correct responses, to get a measure of the real number of pairs of words that the subject knows.

TABLE 36

Per cent of errors in Test I and Test II

	ANTONYMS	NUMBER OF CASES	PER CENT WRONG IN TEST I	PER CENT WRONG IN TEST II	DIFFERENCE BETWEEN TEST I AND TEST II
1	masculine-feminine	150	4	0	−4
2	acquire-lose	150	1	2	1
8	anxiety-nonchalance	139	6	14	8
11	merriment-gloom	140	1	1	0
16	accountable-irresponsible	148	9	9	0
18	immune-susceptible	146	14	12	−2
21	momentous-immaterial	141	12	34	12
22	specific-general	149	14	9	−5
26	waste-conserve	148	2	1	−1
28	any-none	145	7	15	8
30	defile-purify	144	6	10	4
31	ghost-substance	141	6	31	25
39	accumulate-dissipate	139	4	19	15
40	obedient-refractory	133	16	25	9
46	hasty-circumspect	116	22	48	26
49	amiable-surly	122	3	9	6
53	suave-brusque	108	12	13	1
55	superfluous-essential	112	6	12	6
58	vilify-praise	84	17	21	4
60	previous-succeeding	83	24	11	−13
67	contradict-corroborate	62	8	9	1
69	dignify-degrade	60	2	12	10
71	gaunt-buxom	51	16	20	4
74	martial-civil	45	4	4	0
76	vesper-matin	35	40	31	−9
78	even-protuberant	30	13	23	10
41	moderate-violent	135	3	15	12
Average		113.5	9.70	15.1	4.7

One other comparison remains to be discussed. If x is made to equal the sum of errors in both Test I and Test II, then $\frac{8}{7}\,x$ will equal the number of errors in Test I and $\frac{4}{7}\,x$, the number of errors in Test II. (These proportions are derived on the assumption that

has been made throughout all of this discussion, *i.e.*, of the total number of responses which are guessed at one-half will be judged incorrectly and one-half correctly, if there are only two possible responses; and two-thirds will be incorrectly judged and one-third correctly, if there are three possible responses, only one of which is correct.) Of the errors in Test I, or rather, of the pairs of words checked incorrectly in Test I, $\frac{3}{8}$ or $\frac{3}{7}$ x will be checked incorrectly in Test II also; $\frac{1}{3}$ or $\frac{1}{7}$ x will be checked identically the same in both tests.

The errors were actually found to be distributed as follows:

$\frac{8}{7}$ x, total number of errors in Test I	717
$\frac{4}{7}$ x, total number of errors in Test II	992
x, total errors in Tests I and II	1709
$\frac{3}{7}$ x, errors common to both tests, but not necessarily identical	338
$\frac{1}{7}$ x, errors common to both tests and identical	152

Actually, therefore, 19.7 per cent of the total number of errors in both tests were common to both tests, as compared with $\frac{2}{7}$ or 28.5 per cent which would have been expected by chance. Eight and eighty-nine hundredths per cent of the total errors are identical in the two tests, whereas, $\frac{1}{7}$ or 14.28 per cent might have been expected. This is additional evidence that the inclusion of a "neither" group in a synonym-antonym test results not merely in reducing the possibility of making correct responses by chance, but rather, in changing the whole nature of the test. Some of the synonyms and antonyms are actually judged more correctly when the judgment has to take a "neither" category into account. Others are judged less accurately.

Correlations of the two forms of synonym-antonym test with grades

An attempt has been made to relate the two types of synonym-antonym test to some criterion. Since the coefficients of correlation between the score in Test I and the number of correct responses in Test II was 0.770 which indicates that the two tests do not measure identical abilities, it was thought that some clue as to the nature of the difference might be gained from a comparison of their correlations with grades in college subjects. The

coefficients are reported in table 37. There is but small difference between the coefficients, and the coefficients for the second type of test are slightly lower except in the case of the correlation with mathematics grades. It must be remembered that the group represented is selected, not including any engineering students. The fact that the difference between the coefficients is less for the correlation with mathematics, together with the fact that the group is largely a non-science group, suggests that the second type of test may be more predictive of the ability or abilities involved in mathematics and less specialized as an English test. This suggestion, however, is worth but little until a group of engineering students has been tested.

TABLE 37

Coefficients of correlation between average grades and grades in college subjects of the first semester, and two forms of synonym-antonym test. The subjects are members of an English Composition class

	NUMBER OF CASES	TEST I, S.O.	TEST II, S.O.N.
Average grades, first semester	142	0.496 ± 0.04	0.459 ± 0.04
English Composition	138	0.497 ± 0.04	0.425 ± 0.04
Mathematics	113	0.410 ± 0.05	0.414 ± 0.05
French	69	0.554 ± 0.05	0.418 ± 0.06
History	58	0.256 ± 0.08	0.201 ± 0.08

SUMMARY

An analysis of the responses to the different types of stimulus words in two forms of synonym-anatonym test has shown that the inclusion of words that are "neither" similar nor opposite in meaning, with synonyms and antonyms changes the nature of the test. Whereas, theoretically, the chance factor has been reduced, although not eliminated, it has been demonstrated that the presence of the "neither" words complicates the responses, making them more accurate for some pairs of words and more inaccurate for others, so that the chance factor becomes unmeasurable. This fact together with the fact that the errors for specific pairs of words seem to be conditioned more by suggested relationships than by uninfluenced random checking, indicates that it is not possible to assume that a definite proportion of the errors is to be substracted from the correct responses in order to eliminate the chance factor.

CHAPTER IV

Conclusions

All of the tests included in this study are conceded to be verbal in some degree, but they do not have an equal language factor in common. We have considered as language tests those tests that involve an ability to understand abstract verbal concepts, that is, an ability to understand and use verbal language with fine precision. The extent to which such an ability is brought into play or measured determines whether a given test is predominately a language test or not. For example, we have considered the synonym-antonym test in which a factor of primary importance is a knowledge of the exact meanings of words as a language test, whereas tests such as arithmetical reasoning in which the ability to understand actual meanings of words is but a minor factor are not considered language tests.

Four lines of evidence contribute to our knowledge of the extent to which language factors function in specific tests and groups of tests: (*a*) an examination of the test material; (*b*) intercorrelations of tests; (*c*) the relationship to college grades in subjects of distinctly different types as English composition and mathematics; (*d*) differences in performance in tests of groups differing in interests and, accordingly, in course of study, such as arts and sciences and engineering students.

Although no two tests have been found to measure identical abilities, there is more in common between a group of tests, synonym-antonym, opposites, definitions and sentence completion, than between any one of the tests and the ability involved in arithmetical reasoning. These tests have a common language factor in that they all involve the ability to understand and use relatively difficult abstract words. That the highest degree of relationship between any of the tests exists between the synonym-antonym and vocabulary tests argues for the significance of a vocabulary element in the synonym-antonym test. The completion test has relatively more in common with an arithmetical reasoning test than have any of the other tests mentioned above,

which indicates the presence in the test of some potent factor other than a language factor. The information test, on the other hand, has more in common with such tests as synonym-antonym and completion than with arithmetical reasoning.

If grades in English composition are taken as a criterion of language ability as opposed to the grades in mathematics taken as a criterion of mathematical ability, it is found that a group of tests, including the synonym-antonym, opposites and definitions tests especially, and probably analogies, information, syllogisms and reading are measures specifically of language ability as opposed to mathematical ability, whereas the completion test is more nearly equally related to each of the two criteria. Arithmetical reasoning tests are, as could have been predicted, better measures of mathematical ability than of language ability.

Furthermore, since these language tests are on the whole more predictive of success in arts and sciences than in engineering courses, and since the curricula of arts and sciences students are in general more heavily weighted with non-mathematical or language courses than those of the engineers, we have further evidence of a language factor in these tests. The arithmetical reasoning tests are more predictive of the ability of engineers, that is, largely of mathematical ability.

Psychological examinations heavily weighted with tests that have been shown to be more predictive of language ability as involved in English composition and other non-mathematical and non-science courses than of mathematical ability, are less predictive of the work of students interested in engineering, involving as it does both mathematics and science, than of the work of arts and sciences students, who on the average are more interested in the non-mathematical and non-science courses.

Tests for predicting language ability as opposed to mathematical ability are more successful at present than those that predict mathematical ability. This suggests the need of devising and perfecting tests to measure more accurately mathematical ability.

The synonym-antonym test, more than any other single test, has proved to be specifically a test of the ability involved in non-mathematical subjects such as English composition. It has been found that ability to understand the meanings of the stimulus

words is a primary factor in this test but the perception of the relation between words is also a factor. The fact that there is a positive though small degree of correlation with mathematical ability as involved in arithmetical reasoning tests and in mathematics courses, indicates that success in the test is dependent also upon these additional factors.

BIBLIOGRAPHY

(1) Baum, H., Litchfield, M., and Washburn, M. F.: The results of certain standard mental tests as related to the academic record of college seniors. Amer. Jour. Psychol., 1919, xxx, 307–310.

(2) Bell, J. C.: A detailed study of Whipple's range of information test. Jour. Educ. Psychol., 1917, viii, 475–482.

(3) Bell, J. C.: Mental tests and college freshmen. Jour. Educ. Psychol., 1916, vii, 381–399.

(4) Bickersteth, M. E.: The application of mental tests to children of various ages. Brit. Jour. Psychol., 1917–19, ix, 23–73.

(5) Binet, A., and Simon, Th.: The development of intelligence in children. Trans. by Elizabeth Kite. Publications of the Training School at Vineland, N. J., No. 11, 1916.

(6) Bonser, F. G.: The reasoning ability of children of the fourth, fifth and sixth school grades. Colum. Univ. Contrib. Educ., No. 37, 1906.

(7) Brandenburg, G. C.: Psychological aspects of language. Jour. Educ. Psychol., 1918, ix, 313–332.

(8) Briggs, T. H.: Formal English grammar as a discipline. Teachers College Record, 1913, xiv, 215–343.

(9) Brigham, C. C.: A study of American intelligence. Princeton University Press, 1923.

(10) Brown, W.: Some experimental results in the correlation of mental abilities. Brit. Jour. Psychol., 1910, iii, 296–322.

(11) Burt, C.: Experimental tests of higher mental processes and their relation to general intelligence. Jour. Exper. Ped., 1911, i, 93–112.

(12) Carothers, F. E.: Psychological examinations of college students. Archives Psychol., No. 40, 1921.

(13) Carpenter, D. F.: Mental age tests. Jour. Educ. Psychol., 1913, iv, 538–544.

(14) Chapman, J. C., and Dale, A. B.: A further criterion for the selection of mental test elements. Jour. Educ. Psychol., 1922, xiii, 267–276.

(15) Cohn, J., and Dieffenbacher, J.: Untersuchungen über Geschlects-, Alters-, und Begabungs-Unterschiede bei Schulern. Beihefte zur Zeit. Angew. Psych., 1911, ii, 214.

(16) Colvin, S. S.: Psychological tests at Brown University. Sch. and Soc., 1919, x, 27–30.

(17) Decroly et Degand: La mesure de l'intelligence chez des enfants normaux. Archives de Psychol., 1910, ix, 81–108.

(18) Ebbinghaus, H.: Ueber eine neue Methode zur Prufung geistiger Fahigkeiten und ihre Anwendung bei Schulkindern. Zeit. fur Psych., 1897, xiii, 401–459.

(19) Goddard, H. H.: The Binet-Simon measuring scale for intelligence, Rev. Ed. 1911. Depart. Psych. Research, The Training School, Vineland, N. J.
(20) Greene, H. A.: A standardization of certain opposites tests. Jour. Educ. Psychol., 1918, ix, 559.
(21) Henmon, V. A. C.: An experimental study of the value of word study. Jour. Educ. Psychol., 1921, xii, 98–102.
(22) Iowa Entrance Examination. Part III. High School content examination. Devised by G. M. Ruch, 1923.
(23) James, B. B.: Mutual correlation of intelligence, scholarship and vocabulary. Sch. and Soc., 1919, ix, 427.
(24) James, W.: On a very prevalent abuse of abstraction. Pop. Science Monthly, 1909, lxxiv, 485–493.
(25) Jordan, A. M.: Correlations of four intelligence tests with grades. Jour. Educ. Psychol., 1922, xiii, 419–429.
(26) Jordan, A. M.: The validation of intelligence tests. Jour. Educ. Psychol., 1923, xiv, 348–366; 414–428.
(27) Kelley, T. L.: Distinctive ability. Sch. and Soc., 1923, xviii, 424–428.
(28) Kelley, T. L.: Reliability of test scores. Jour. Educ. Research, 1921, iii, 370–379.
(29) King, I., and Gold, H.: A tentative standardization of certain "opposites tests." Jour. Educ. Psychol., 1916, vii, 459–482.
(30) King, I., and M'Crory, J. L.: Freshmen tests at the State University of Iowa. Jour. Educ. Psychol., 1918, ix, 32–46.
(31) Kirkpatrick, E. A.: A vocabulary test. Pop. Science Monthly, 1907, lxx, 157–164.
(32) Kitson, H. D.: The scientific study of college students. Psychol. Monog., 1917, xxiii, No. 89.
(33) Kuhlmann, F.: A Handbook of Mental Tests. (A further revision and extension of the Binet-Simon Scale.) Baltimore, Warwick and York, 1922.
(34) McCall, W. A.: Correlation of some educational and psychological measurements. Colum. Univ. Contrib. Educ., No. 79, 1916.
(35) Means, M. H.: A tentative standardization of a hard opposites test. Psychol. Monog., 1921, xxv, 137, pp. 65.
(36) Norsworthy, N.: The Psychology of Mentally Deficient Children. New York, 1906.
(37) Otis, A. S.: Otis Group Intelligence Scale. Forms A and B. New York, World Book Co., 1919.
(38) Pintner, R.: The measurement of progress in language ability. Jour. Educ. Psychol., 1918, ix, 270–277.
(39) Pintner, R., and Paterson, D. G.: A measurement of the language ability of deaf children. Psychol. Rev., 1916, xxiii, 413–436.
(40) Pintner, R., and Renshaw, S.: A standardization and weighting of two hundred analogies. Jour. App. Psychol., 1920, iv, 263–273.

(41) Pyle, W. H.: The Examination of School Children. New York, 1913.
(42) Pyle, W. H.: The Examination of School Children. A manual for the mental and physical examination of school children. (Rev.) Univ. Missouri Bull., 1920, xxi, No. 12.
(43) Simpson, B. R.: Correlations of mental abilities. Colum. Univ. Contrib. Educ. No. 53, 1912.
(44) Smith, L. L.: Whipple's range of information test. Psychol. Rev., xx, 517–518.
(45) Squire, C. R.: Graded mental tests. Jour. Educ. Psychol., 1912, iii, 363–375; 430–443.
(46) Symposium: Intelligence and its measurement. Jour. Educ. Psychol., 1921, xii, 123–147; 195–216; 271–275.
(47) Terman, L. M.: Genius and stupidity. Ped. Sem., 1906, xiii, 307–373.
(48) Terman, L. M.: Intelligence tests in colleges and universities. Sch. and Soc., 1921, xiii, 481–494.
(49) Terman, L. M.: Some data on the Binet test of naming words. Jour. Educ. Psychol., 1919, x, 29–35.
(50) Terman, L. M.: Terman Group Test of Mental Ability. New York, World Book Co., 1920.
(51) Terman, L. M.: The Measurement of Intelligence. New York, Houghton Mifflin Co., 1916, pp. 362.
(52) Terman, L. M.: The vocabulary test as a measure of intelligence. Jour. Educ. Psychol., 1918, ix, 452–466.
(53) Terman, L. M., and Childs, H. G.: A tentative revision and extension of the Binet-Simon measuring scale of intelligence. Jour. Educ. Psychol., 1912, iii, 61–74; 133–143; 198–208; 277–289.
(54) Thorndike, E. L.: Measurement of Twins. Colum. Univ. Contrib. Philos. and Psychol., 1905, xiii, No. 3.
(55) Thurstone, L. L.: A cycle-omnibus intelligence test for college students. Jour. Educ. Research, 1921, iv, 265–278.
(56) Tolman, E. C.: English and mathematical abilities of a group of college students. Jour. Educ. Psychol., 1919, x, 95–103.
(57) Toops, H. A.: Eliminating the pitfalls in solving correlation: a printed correlation form. Jour. Exper. Psychol., 1921, iv, 434–446.
(58) Trabue, M. R.: Completion-test language scales. Contrib. Educ., No. 77. N. Y. Teachers College, 1916.
(59) Twenty-first Yearbook of the National Society for the Study of Education: Intelligence Tests and Their Use. Whipple, G. M., Ed., Bloomington, Ill., Public School Publishing Co., 1922.
(60) Uhl, W. L.: Mentality tests for college freshmen. Jour. Educ. Psychol., 1919, x, 13–28.
(61) Van Wagenen, M. J.: Graded opposites and analogies tests. Jour. Educ. Psychol., 1920, xi, 240–263.
(62) Van Wagenen, M. J., and Kelly, F. E.: Language abilities and their relations to college marks. Jour. Educ. Psychol., 1920, xi, 459–473.

(63) WHIPPLE, G. M.: A range of information test. Psychol. Rev., 1909, xvi, 347–351.
(64) WHIPPLE, G. M.: Endowment, maturity and training as factors in intelligence scores. Scientific Monthly, 1924, xviii, 496–507.
(65) WHIPPLE, G. M.: Manual of mental and physical tests. Part II. Complex processes. Baltimore, Warwick and York, 1915.
(66) WHIPPLE, G. M.: Vocabulary and word building tests. Psychol. Rev., 1908, xv, 94–105.
(67) WHIPPLE, G. M., HENRY, MANUEL, COY: Classes for Gifted Children. Bloomington, Ill., Public School Publishing Co., 1919.
(68) WIERSMA, E.: Die Ebbinghaus'che Combinationsmethode. Zeit. fur. Psych., 1902, xxx, 196–222.
(69) WOOD, B. D.: Measurement in higher education. New York, Teachers College, Columbia University, 1923.
(70) WOODWORTH, R. S., AND WELLS, F. L.: Association tests. Psychol. Monog., 1911, xiii.
(71) WOOLLEY, H. T., AND FISCHER, C. R.: Mental and physical measurements of working children. Psychol. Monog., 1914, xviii, 77, 213–341.
(72) WYATT, S.: The quantitative investigation of higher mental processes. Brit. Jour. Psychol., 1913, vi, 109–133.
(73) YERKES, R. M., BRIDGES, J. W., AND HARDWICK, R. S.: A Point Scale for Measuring Mental Ability. Baltimore, Warwick and York, 1915.
(74) YERKES, R. M.: Psychological examining in the United States Army. Memoirs XV, National Acad. Sciences, Washington, 1921.
(75) YULE, G. U.: An Introduction to the Theory of Statistics. 6th ed. London, Charles Griffin and Co., Ltd., 1922.

MONOGRAPHS ISSUED

I. An Analysis of Language Factors in Intelligence Tests. *Dorothy W. Scago*

II. Intelligence and Immigration. *Clifford Kirkpatrick*

III. A Psychological Study of Immigrant Children at Ellis Island. *Bertha M. Boody*

Other announcements will follow